Aviation Security Screening

Competency Assessment from Different Perspectives

Thesis
presented to the Faculty of Arts
of
the University of Zurich
for the degree of Doctor of Philosophy

by
Saskia M. Koller
of Zürich ZH

Accepted in the autumn semester 2008 on the recommendation of
Prof. Dr. Wolfgang Marx and Prof. Dr. Damian Läge

ETH Press
Zurich 2008

Danksagung

An dieser Stelle möchte ich allen Personen danken, die mich während der Zeit meiner Promotion begleitet und unterstützt haben und zum Gelingen dieser Arbeit beigetragen haben.

Prof. Dr. Wolfgang Marx möchte ich danken für die Unterstützung und Förderung schon während meines Studiums. Ein grosser Dank geht an meinen Betreuer, Prof. Dr. Adrian Schwaninger, der mich mit seinem grossen Fachwissen schon während des Studiums für das Themengebiet begeistern konnte. Er brachte mir professionelles wissenschaftliches Arbeiten in Verbindung mit angewandter Forschung bei und hat mich stets gefördert und gefordert. Besten Dank auch an Prof. Dr. Damian Läge für die Begutachtung der Arbeit.

Die vorliegende Arbeit konnte nur durch die Unterstützung und in Zusammenarbeit mit verschiedenen Partnern entstehen.

Ein herzliches Dankeschön geht an alle Personen der Visual Cognition Research Group, die mich in irgendeiner Art unterstützt haben. Kapitel 1 entstand unter Mithilfe von Anton Bolfing, der mich während der ganzen Zeit mit seinem grossen Wissen unterstützt hat. Für die Unterstützung zur Entstehung von Kapitel 2 möchte ich Diana Hardmeier und Stefan Michel danken, auch sie haben mich in inspirierenden Gesprächen und mit viel Fachwissen oft gefördert. Stefan Michel half auch mit in Kapitel 4, Diana Hardmeier gab wertvollen Input für die Studien in Kapitel 6 und 7. Kapitel 3 entstand mit der Hilfe und durch Inspiration von Prof. Dr. Colin Drury. Auch unseren Programmierern möchte ich herzlich danken für die grossartige Unterstützung, vor allem Markus Ruh für die Mitarbeit in Kapitel 4 und auch

ganz speziell Jonas Sourlier für die ganze Programmierung der Studie in Kapitel 5. Jiwen Li möchte ich ebenfalls für die Mithilfe in Kapitel 5 danken, welche eine interdisziplinäre Studie zwischen Psychologie und Informatik ermöglicht hat. Schliesslich danke ich Anita Sennhauser für die Mithilfe der Studie in Kapitel 7. Nicht zu vergessen sind alle Versuchsteilnehmer, nämlich die Luftsicherheitsangestellten an diversen Flughäfen, die unsere Forschungsprojekte überhaupt erst ermöglichen - vielen Dank.

Ganz grosse Dankbarkeit gebührt meiner Familie und meinen Freunden, die mir den Weg geebnet, mich ständig begleitet und kompromisslos unterstützt haben. Tausend Dank für Euer Verständnis in strengen Zeiten. In grosser Liebe danke ich Roger für seine bedingungslose und selbstverständliche Unterstützung während meiner gesamten Studienzeit und ganz speziell während des intensiven Schlussspurts.

Saskia M. Koller

Zürich, im August 2008

Contents

Summary

This thesis deals with the human factor in aviation security screening. In the last decades the threat level of terroristic attacks has increased dramatically. This also affects aviation. By and by different security measures were introduced and in the 1970s airports began implementing the screening of passengers and their baggage. Thereby it should be prevented that any object can be brought aboard an aircraft that could be used to do harm. This task is being executed by aviation security screening officers.

In the course of the development in recent years it has been realized more and more that the position of the screening officers has to be strengthened. They are - as human factor - one of the weakest links in the whole screening procedure. Still they have to take the final decision whether a passenger and his bag are clear and thus allowed to board the aircraft. Technologies are advanced constantly, but also screening officers should receive attention. The aim is to turn the human factor into the strongest link in the screening procedure. Therefore, investments have to be made into careful selection and specific training of the screening officers in order to lay the foundations for reaching and maintaining high standards and quality. One essential aspect to even make this possible is the assessment of the screening officers' competencies. In this thesis the focus lies on the X-ray screening task within the screening procedure. Different aspects of aviation security X-ray screening are attended to.

The first part of this thesis deals with competency assessment and training. Chapter 1 presents a detailed introduction into assessment and certification of X-ray image

interpretation competency of aviation security screening officers. Certification is a very important means in the attempt to introduce and maintain high standards. It is disclosed what is important for serious and professional competency assessment and certification. The study in Chapter 2 investigated the effect of computer-based training for X-ray image interpretation competency. Large effects were found in that the X-ray image interpretation competency, assessed with an appropriate test, could be increased significantly when screening officers received training. It could also be shown a transfer of knowledge due to training. This is an essential finding and one headstone in the intention to strengthen the human factor in the process of aviation security screening. Chapter 3 takes a different approach on the investigation of training effects. Here, the time used to interpret an X-ray image of a bag and searching it for a threat object is analyzed. By applying a special formula, portions of this reaction time can be allocated to different processes involved in the interpretation of an image. It was assumed that on the one hand the image is searched visually, that is, scanned. The portion of time needed for this process is called search time. On the other hand, time is needed to take a decision. Objects first have to be detected in the image and then they have to be identified and a decision has to be taken if it is a threat object or a harmless object. This is the decision time. Chapter 3 investigated the effect of training on these two interpretation processes by analyzing the search time and the decision time separately. The effect of training cropped up in both processes, in decision time even more than in search time. This means that not only the scanning process can be improved by training, but the specific learning of threat objects decreases the decision time. That is, threat objects are detected faster and the interpretation and recognition becomes faster as well.

In the second part of this thesis potential decision aids for the X-ray image interpretation task of aviation security screening officers are investigated. The study in Chapter 4 explored the image enhancement feature which is integrated in most of the X-ray machines. Image enhancements are algorithms for the different display of an X-ray image. Modern X-ray machines color-code the different materials being scanned. For example, this artificial coloring displays metallic material in blue and organic material in orange. Of course there are many different shadings and more

colors. Image enhancements are supposed to facilitate the interpretation of an X-ray image in that they provide the possibility to suppress certain image information and bring out other. For example, the Metallic Only filter displays only the blue-colored metallic material, the Organic Only filter only displays organic material and suppresses all other material, etc. These image enhancements are provided by the X-ray machine manufacturers without having been investigated regarding their effectiveness. This was the aim of the study in Chapter 4. The present study shows that X-ray image interpretation could not be improved with these image enhancement filters. Chapter 5 applies a concept on X-ray image interpretation that originated in marketing. Marketing, dealing with consumers and their behavior among others, devides the products on a market into different sets. Of all products that are available on the market, the consideration set includes those products a consumers takes into account when choosing the product to be bought. For our study, the concept of the consideration set is adapted to X-ray image interpretation. As a decision aid for aviation security screening officers, when interpreting an X-ray image of a passenger bag, consideration information was provided. Consideration information were X-ray image patches that were presented simultaneously to the bag. The results of this study indicate that there is no benefit of consideration information on detection performance of threat objects in X-ray images of passenger bags.

The third part of the thesis is more of statistical and test theoretical nature. Chapter 6 deals with the analysis of tests which are used for certification of aviation security screening officers. Different ways of analysis were applied and their effect on test difficulty and reliability were investigated. There could actually be found a rather large impact of analysis method which confirms the need to take a closer look at the scoring method that is used for assessing competencies of employees. The results provide a good basis for the decision which analysis method is appropriate to use for certification purposes in aviation security screening. In Chapter 7 the Angoff method is applied to X-ray image tests. The Angoff method is a widely used method to establish a pass mark for a test with a criterion-referenced approach instead of a normative approach. The first aim was to investigate if the Angoff method is appropriate to use for image tests since in the literature it is mainly applied to theoretical

exams, mostly multiple-choice exams. Reliability analyses confirmed the applicability of the Angoff method for establishing a pass mark for image tests. A passmark was established for an X-ray image test which is used for certification of aviation security screening officers. This criterion-referenced pass mark is compared to the normative passmark which has been applied until now. The results are comparable.

The main contribution of this thesis are the new approaches that were applied to aviation security X-ray image interpretation competency assessment - besides the formulation of important basic prerequisites for fundamental headstones in the ambitious intention to turn the human operator into the strongest link within the aviation security screening process.

Part I

Competency Assessment and Training

1

Assessment and Certification of X-Ray Image Interpretation Competency of Aviation Security Screeners

1.1 Competency Assessment in Aviation Security Screening

1.1.1 Introduction

In response to the increased risk of terrorist attacks, large investments have been made in recent years in aviation security technology. However, the best equipment is of limited value if the people who operate it are not selected and trained appropriately to perform their tasks effectively and accurately. Latterly, the relevance of human factors has increasingly been recognized. One important aspect of human factors is the competency of the personnel who conducts security screening at airports (aviation security screeners) and assessment of their competency. Competency assessment is important in maintaining the workforce certification process. The main aim of certification procedures is to ensure that adequate standards in aviation security are consistently and reliably achieved. Certification of aviation security screeners can be considered as providing quality control over the screening process. Using certification tests, important information on strengths and weaknesses in aviation security procedures in general as well as on each individual screener can be obtained. As a consequence, certification can also be a valuable basis for qualifying personnel, measuring training effectiveness, improving training procedures and increasing motivation. In short, certification and competency assessment can be important instruments in improving aviation security. The implementation of competency assessment procedures has several challenges. First, what should be assessed has to be identified. Then, it should be considered how procedures for certification

of different competencies can be implemented. Another important challenge is international standardization, since several countries, organizations and even companies are independently developing their own certification or quality control systems. The following international documents refer to certification and competency assessment of aviation security staff:

- EU Regulation 2320/2002
- ICAO Annex 17, 3.4.3[1]
- ICAO-Manual on Human Factors in Civil Aviation Security Operations (Doc 9808)[2]
- ICAO Human Factors Training Manual (Doc 9683), Part 1, Chapter 4, and Appendix 6, Appendix 32[3]
- ICAO Security Manual for Safeguarding Civil Aviation Against Acts of Unlawful Interference, Doc 8973, Chapter 4, I-4-45[4]
- ECAC Doc 30, Chapter 12, and Annex IV-12A[5]

ECAC Doc 30 by the European Civil Aviation Conference specifies three elements for *initial* certification of aviation security screeners:

- an X-ray image interpretation exam
- a theoretical exam
- a practical exam

Periodic certification should contain a theoretical exam and an X-ray image interpretation exam. Practical exams can be conducted if considered necessary. This chapter covers the first element, that is, how to examine competency in X-ray image

[1] ICAO Annex 17, 3.4.3 (Each Contracting State shall ensure that the persons carrying out screening operations are certified according to the requirements of the national civil aviation security programme)
[2] ICAO Manual on Human Factors in Civil Aviation Security Operations (Doc 9808)
[3] ICAO Human Factors Training Manual (Doc 9683), Part 1, Chapter 4, and in Appendix 6 - Guidance on recruitment, selection, training, and certification of aviation security staff and Appendix 32 - Guidance on the use of threat image projection
[4] ICAO Security Manual for Safeguarding Civil Aviation Against Acts of Unlawful Interference, Doc 8973, Chapter 4, I-4-45 (Recruitment, selection, training and certification of security staff)
[5] ECAC Doc 30 Annex IV-12A Certification criteria for screeners and ECAC Doc 30, Chapter 12, 12.2.3 Certification of security staff 1.1.10.3

interpretation. First, human factors best practice guidance for assessing the X-ray image interpretation competency of aviation security screeners is provided. Three different possibilities are discussed, which can serve to measure competency in X-ray image interpretation: covert testing, threat image projection (TIP), and computer-based image tests. Second, on-the-job assessment of screener competency using TIP is discussed. Third, an example of a reliable, valid, and standardized computer-based test is presented which is now used at more than 100 airports worldwide to measure competency in X-ray image interpretation and also for certification purposes, the X-Ray Competency Assessment Test (X-Ray CAT). Fourth, the application of this test in an EU-funded project (the VIA Project) at several European airports is presented.

1.1.2 Requirements for assessing competency

One of the most important tasks of an aviation security screener is the interpretation of X-ray images of passenger bags and the identification of prohibited items within these bags. Hit rates, false alarm rates, and the time used to visually inspect an X-ray image of a passenger bag are important measures that can be used to assess the effectiveness of screeners at this task. A hit refers to detecting prohibited items in an X-ray image of a passenger bag. The hit rate refers to the percentage of all X-ray images of bags containing a prohibited item that are correctly judged as being NOT OK. If a prohibited item is reported in an X-ray image of a bag that does not contain such an item, this counts as a false alarm. The false alarm rate refers to the percentage of all harmless bags (i.e., bags not containing any prohibited items) that are judged as containing a prohibited item by a screener. The time taken to process each bag is also important as it helps in determining throughput rates and can indicate response confidence. The results of an X-ray image interpretation test provide very important information for civil aviation authorities, aviation security institutions, and companies. Moreover, failing a test can have serious consequences, depending on the regulations of the appropriate authority. Therefore, it is essential that a test is fair, reliable, valid, and standardized. In the last 50 years, scientific criteria have been developed that are widely used in psychological testing and psy-

chometrics. These criteria are essential for the development of tests for measuring human performance. A summary of the three most important concepts, namely reliability, validity, and standardization, is now presented. A more detailed introduction to psychological testing and psychometrics can be found in handbooks such as Fishman and Galguera (2003), Kline (2000), or Murphy and Davidshofer (2001).

Reliability

We mean reliability to refer to the "consistency" or "repeatability" of measurements. It is the extent to which the measurements of a test remain consistent over repeated tests of the same participant under identical conditions. If a test yields consistent results of the same measure, it is reliable. If repeated measurements produce different results, the test is not reliable. If, for example, an IQ test yields a score of 90 for an individual today and 125 a week later, it is not reliable. The concept of reliability is illustrated in Figure 1.1. Each point represents an individual person. The x-axis represents the test results in the first measurement and the y-axis represents the scores in the same test of the second measurement. Figures 1.1a-c represent tests of different reliability. The test in Figure 1.1a is not reliable. The score a participant achieved in the first measurement does not correspond at all with the test score of the second measurement. The reliability coefficient can be calculated by the correlation between the two measurements. In Figure 1.1a the correlation is near zero, that is, $r = 0.05$ (the theoretical maximum is 1). The test in Figure 1.1b is somewhat more reliable. The correlation between the two measurements is 0.50. Figure 1.1c shows a highly reliable test with a correlation of 0.95.

The reliability of a test may be estimated by a variety of methods. When the same test is repeated (usually after a time interval during which job performance is assumed not to have changed), the correlation between the scores achieved on the two measurement dates can be calculated. This measure is called *test-retest reliability*. A more common method is to calculate the *split-half reliability*. With this method, the test is divided into two halves. The whole test is administered to a sample of participants and the total score for each half of the test is calculated. The split-

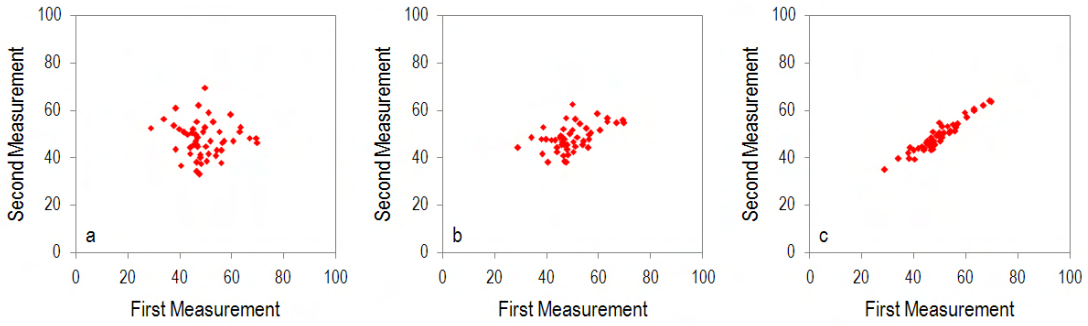


Fig. 1.1. Illustration of different correlation coefficients. a: $r = 0.05$, b: $r = 0.50$, c: $r = 0.95$.

half reliability is the correlation between the test scores obtained in each half. In the alternate forms method, two tests are created that are equivalent in terms of content, response processes, and statistical characteristics. Using this method, participants take both tests and the correlation between the two scores is calculated (*alternate forms reliability*). Reliability can also be a measure of a test's internal consistency. Using this method, the reliability of the test is judged by estimating how well the items that reflect the same construct or ability yield similar results. The most common index for estimating the internal reliability is Cronbach's alpha. Cronbach's alpha is often interpreted as the mean of all possible split-half estimates. Another internal consistency measure is KR 20 (for details see standard text books on psychometrics such as for example Fishman & Galguera, 2003; Kline, 2000; Murphy & Davidshofer, 2001). Acceptable tests usually have reliability coefficients between 0.7 and 1.0. Correlations exceeding 0.9 are not often achieved. For individual performance to be measured reliably, correlation coefficients of at least 0.75 and Cronbach's alpha of at least 0.85 are recommended. These numbers represent the minimum values. In the scientific literature, the suggested values are often higher.

Validity

Validity indicates whether a test is able to measure what it is intended to measure. For example, the hit rate alone is not a valid measure of detection performance in terms of discriminability (or sensitivity), because a high hit rate can also be achieved by judging most bags as containing prohibited items. In order to measure detection

performance in terms of discriminability (or sensitivity), the false alarm rate must be considered, too (for different detection measures see Macmillan & Creelman, 1991, and Hofer & Schwaninger, 2004). As with reliability, there are also different types of validity. The term *face validity* refers to whether a test appears to measure what it claims to measure. A test should reflect the relevant operational conditions. For example, if a test for measuring competency in X-ray image interpretation contains a representative sample of X-ray images of bags and screeners have to decide whether the depicted bags contain a prohibited item, it is *face valid*. *Concurrent validity* refers to whether a test can distinguish between groups that it should be able to distinguish between (e.g., between trained and untrained screeners). In order to establish *convergent validity*, it has to be shown that measures that should be related are indeed related. If, for example, threat image projection (TIP, i.e., the insertion of fictional threat items into X-ray images of passenger bags) measures the same competencies as a computer-based offline test, one would expect a high correlation between TIP performance data and the computer-based test scores. Another validity measure is called predictive validity. In *predictive validity*, the test's ability to predict something it should be able to predict is assessed. For example, a good test for pre-employment assessment would be able to predict on-the-job X-ray screening detection performance. *Content validity* refers to whether the content of a test is representative of the content of the relevant task. For example, a test for assessing whether screeners have acquired the competency to detect different threat items in X-ray images of passenger bags should contain X-ray images of bags with different categories of prohibited items, according to an internationally accepted prohibited items list.

Standardization and development of population norms

The third important aspect for judging the quality of a test is standardization. This involves administering the test to a representative group of people in order to establish norms (a normative group). When an individual takes the test, it can then be determined how far above or below the average her or his score is, relative to the normative group. It is important to know how the normative group was

selected, though. For instance, for the standardization of a test used to evaluate the detection performance of screeners, a meaningful normative group of a large and representative sample of screeners (at least 200 males and 200 females) should be tested. In summary, competency assessment of X-ray image interpretation needs to be based on tests that are reliable, valid, and standardized. However, it is also important to consider test difficulty, particularly if results from different tests are compared with each other. Although two tests can have similar properties in terms of reliability, an easy test may not adequately assess the *level* of competency needed for the X-ray screening job.

1.1.3 Assessment of X-ray image interpretation competency

Currently, there are several methods used to assess X-ray image interpretation competency: covert testing (infiltration testing), threat image projection (TIP), and computer-based image tests.

Covert testing

Covert testing, as the exclusive basis for individual assessment of X-ray image interpretation competency, is only acceptable if the requirements of reliability, validity, and standardization are fulfilled. For covert testing to achieve these requirements, a significant number of tests of the same screener is necessary in order to assess competency reliably. More research is needed to address this issue and it should be noted that this chapter does not apply to principles and requirements for covert testing used to verify compliance with regulatory requirements.

Threat image projection (TIP)

Screener competency can also be assessed using TIP data if certain requirements are met. In cabin baggage screening, TIP is the projection of fictional threat items into X-ray images of passenger bags during the routine baggage screening operation.

This way, the detection performance of a screener can be measured under operational conditions. Using *raw* TIP data alone does not provide a reliable measure of individual screener detection performance. Data needs to be *aggregated* over time in order to have a large enough sample upon which to perform meaningful analyses. In order to achieve reliable, valid, and standardized measurements, several other aspects need to be taken into account as well when analyzing TIP data. One requirement is to use an appropriate TIP library. This should contain a large number of threat items, which represent the prohibited items that need to be detected and which feature a reasonable difficulty level. See the section on reliable measurement of performance using TIP for more information on how to use TIP data for measuring X-ray detection performance of screeners.

Computer-based X-ray image interpretation tests

Computer-based X-ray image interpretation tests constitute a valuable tool for standardized measurements of X-ray image interpretation competency. These tests should consist of X-ray images of passenger bags containing different prohibited objects. The categories of items should reflect the prohibited items list and requirements of the appropriate authority, and it should be ensured that the test content remains up-to-date. The test should also contain harmless bag images, that is, X-ray images of bags that do not contain a prohibited object. For each image, the screeners should indicate whether or not a prohibited object is present. Additionally, the screeners can be requested to identify the prohibited item(s). The image display duration should be comparable to operational conditions. Test conditions should be standardized and comparable for all participants. For example, the brightness and contrast on the monitor should be calibrated and similar for all participants. This applies equally to other monitor settings that could influence detection performance (e.g., the refresh rate). In order to achieve a valid measure of detection performance, not only hit rates but also false alarm rates should be taken into account. An additional or alternative measure would be to count the number of correctly identified prohibited items (in this case, candidates have to indicate where exactly in the bag the threat is located). The test should be reliable, valid, and standardized. Reliabil-

ity should be documented by scientifically accepted reliability estimates (see above, section 1.1.2). If possible, validity measures should also be provided (see above, section 1.1.2). Individual scores should be compared to a norm that is based on a large and representative sample of screeners (see above, section 1.1.2). The probability of detecting a prohibited item depends on the knowledge of a screener as well as on the general difficulty of the item. Image-based factors such as the orientation in which a threat item is depicted in the bag (view difficulty), the degree by which a threat object is superimposed by other objects (superposition), and the number and type of other objects within the bag (bag complexity) influence detection performance substantially (Schwaninger, Hardmeier, & Hofer, 2004; Schwaninger, 2003b). Tests should take these effects into account.

1.1.4 Certification of X-ray image interpretation competency

As indicated above and as specified in ICAO Annex 17, 3.4.3, individuals carrying out screening operations should be certified initially and periodically thereafter. Certification can not only be considered as providing quality control over the screening process, but also as a valuable basis for awarding personnel a qualification, measuring training effectiveness, improving training procedures, and increasing motivation. Certification data provides important information on strengths and weaknesses in aviation security procedures in general as well as on individual screeners. Furthermore, standardized certification can help in achieving international standardization in aviation security. However, this is very challenging, since many countries, organizations, and companies develop their own certification and quality control systems. The present section gives a brief overview of how a certification can be implemented. As mentioned above, certification of screeners should contain a theoretical exam and an X-ray image interpretation exam. For periodic certification, practical exams can be conducted if considered necessary, unlike the initial certification, where practical exams are required. The exams should meet the requirements of high reliability and validity and standardization (see above).

The X-ray image interpretation exam should be adapted to the location in which a screener is employed, that is, cabin baggage screening, hold baggage screening, or cargo screening. Since not every threat object always constitutes a threat during the flight, depending on where aboard the aircraft it is transported, screeners should be certified according to their location. The certification of cabin baggage screeners should be based on cabin baggage X-ray images that contain all kinds of objects that are prohibited from being carried on in cabin baggage (guns, knives, improvised explosive devices, and other prohibited items). Objects that are prohibited from being transported in the cabin of an aircraft do not necessarily pose a threat when transported in the hold or in cargo. Furthermore, different types of bags are transported in the cabin, the hold, and cargo. Usually, small suitcases or bags serve as hand baggage, whereas big suitcases and traveling bags are transported in the hold of the aircraft. The certification of hold baggage screeners should be carried out using X-ray images of hold baggage. Similarly, cargo screeners should be tested using X-ray images of cargo items. Screeners should be kept up-to-date regarding new and emerging threats. In order to verify whether this is consistently achieved, it is recommended that a recurrent certification should be conducted on a periodical basis, typically every 1-2 years. The minimum threshold that should be achieved in the tests in order to pass certification should be defined by the national appropriate authority and should be based on a large and representative sample of screeners (see also the section on standardization for more information on this topic).

1.2 Measurement of Performance on the Job Using Threat Image Projection (TIP)

Threat image projection (TIP) is a function of state-of-the-art X-ray machines that allows the exposure aviation security screeners to artificial but realistic X-ray images during the process of the routine X-ray screening operation at the security checkpoint. For cabin baggage screening (CBS), fictional threat items (FTIs) are digitally projected in random positions into X-ray images of real passenger bags. In hold baggage screening (HBS), combined threat images (CTIs) are displayed on the

monitor. In this case, not only the threat item is projected but an image of a whole bag which may or may not contain a threat item. This is possible if the screeners visually inspecting the hold baggage are physically separated from the passengers and their baggage. If a screener responds correctly by pressing a designated key on the keyboard (the "TIP key") it counts as a hit, which is indicated by a feedback message. If a screener fails to respond to a projected threat within a specified amount of time, a feedback message appears indicating that a projected image was missed. This would count as a miss. Feedback messages also appear if a screener reports a threat although there was no projection of a threat or a CTI. In this case, it could be a real threat. Projecting whole bags in HBS provides not only the opportunity to project threat images (i.e., bags containing a threat item) but also non-threat images (i.e., bags not containing any threat item). This also allows the recording of false true alarms (namely, if a non-threat image was judged as containing a threat) and correct rejections (namely, if a non-threat image was judged as being harmless). TIP data are an interesting source for various purposes like quality control, risk analysis, and assessment of individual screener performance. Unlike the situation in a test setting, individual screener performance can be assessed on the job when using TIP data. However, if used for the measurement of individual screener X-ray detection performance, international standards of testing have to be met, that is, the method needs to be reliable, valid, and standardized (see above). In a study on CBS and HBS TIP, Hofer and Schwaninger (2005) found very low reliability values for CBS TIP data when a small TIP image library of a few hundred FTIs was used. Good reliabilities were found for HBS TIP data when a large TIP image library was available. The authors suggest that a large image library (at least 1000 FTIs) containing a representative sample of items of varying difficulty should be used when TIP is used for individual performance assessment. Also viewpoint difficulty, superposition, and bag complexity may need to be considered. Finally, as mentioned above, data needs to be aggregated over time in order to have a large enough sample upon which to perform meaningful analyses. TIP data should only be used for certification of screeners if the reliability of the data has been proven, for example by showing that

the correlation between TIP scores based on odd days and even days aggregated over several months is higher than .75.

1.3 X-Ray Competency Assessment Test (X-Ray CAT)

1.3.1 Introduction

This section introduces the X-Ray Competency Assessment Test (X-Ray CAT) as an example of a computer-based test that can be used for assessing X-ray image interpretation competency. The CAT has been developed on the basis of scientific findings regarding threat detection in X-ray images of passenger bags (Schwaninger et al., 2004; Schwaninger, 2003b). How well screeners can detect prohibited objects in passenger bags is influenced in two different ways. First, it depends on the screener's knowledge of what objects are prohibited and what they look like in X-ray images. This knowledge is an attribute of the individual screener and can be enhanced by specific training. Second, the probability of detecting a prohibited item in an X-ray image of a passenger bag also depends on image-based factors. These are the orientation of the prohibited item within the bag (view difficulty), the degree by which an object is superimposed by other objects in the bag (superposition), and the number and type of other objects within the bag (bag complexity). Systematic variation or control of the image-based factors is a fundamental property of the test and has to be incorporated in the test development. In the X-Ray CAT, the effects of viewpoint are controlled by using two standardized rotation angles in an 'easy' and a 'difficult' view for each forbidden object. Superposition is controlled in the sense that it is held constant over the two views and as far as possible over all objects. With regard to bag complexity, the bags are chosen in such a way that they are visually comparable in terms of the form and number of objects with which they are packed. The X-Ray CAT contains two sets of objects in which object pairs are similar in shape. This construction not only allows the measurement of any effect of training, that is, if detection performance can be increased by training, but also possible transfer effects. The threat objects of one set can be included in the training.

By measuring detection performance after training using both sets of the test, it can be ascertained whether training also helped in improving the detection of the objects that did not appear during training. Should this be the case, it indicates a transfer of the knowledge gained about the visual appearance of objects used in the training to similar looking objects.

1.3.2 Materials for the test

Stimuli were created from color X-ray images of prohibited items and passenger bags (Figure 1.2 displays an example of the stimuli).

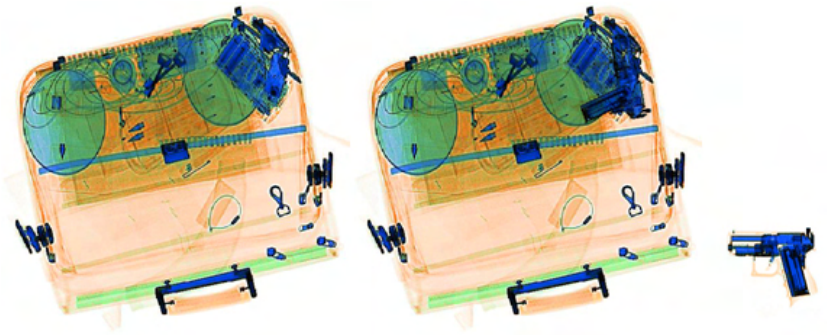


Fig. 1.2. Example images from the X-Ray CAT. Left: harmless bag (non-threat image), right: same bag with a prohibited item at the top right corner (threat image). The prohibited item (gun) is also shown separately at the bottom right.

On the basis of the categorization of current threat image projection systems (Doc 30 of the European Civil Aviation Conference, ECAC), four categories of prohibited items were chosen to be included in the test: guns, improvised explosive devices (IEDs), knives, and other prohibited items (e.g., gas, chemicals, grenades, etc.). The prohibited items were selected and prepared in collaboration with airport security experts to be representative and realistic. Sixteen exemplars are used of each category (eight pairs). Each pair consists of two prohibited items of the same kind that are similar in shape. The pairs were divided into two sets, set A and set B. Furthermore, each object within both sets is used in two standardized viewpoints (see Figure 1.3). The easy viewpoint shows the object in canonical (easily recognizable) perspective (Palmer, Rosch, & Chase,

1981); the difficult viewpoint shows it with an 85 degree horizontal rotation or an 85 degree vertical rotation. In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically and the other half horizontally. The corresponding object of the other set is rotated around the same axis.

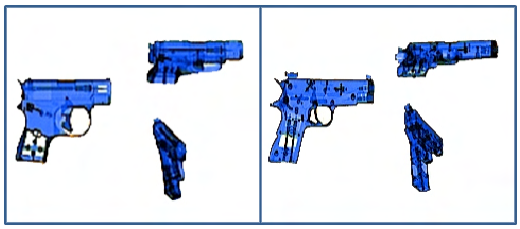


Fig. 1.3. Example of two X-ray images of similar looking threat objects used in the test. Left: a gun of set A. Right: Corresponding gun of set B. Both objects are depicted also in 85 degree horizontal rotation (top) and 85 degree vertical rotation (bottom).

In order to compare how well a prohibited object can be detected relative to its counterpart in the other image set, the two conditions should be comparable in regard to the rotation of the objects, superposition by other objects, and the bag complexity. Furthermore, the superposition should also be the same for both viewpoints of an object. This was achieved using an image-processing tool to combine the threat objects with passenger bags of comparable image complexity, and at the same time, controlling for superposition. This tool calculates the difference in brightness between the pixels of the two superimposed images (threat object and bag) using the following formula for superposition:

$$SP = \frac{\sqrt{\sum [I_{SN}(x,y) - I_N(x,y)]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); ObjectSize: Number of pixels of the prohibited item where R, G, and B are < 253

This equation, developed especially for the preparation of these test images) calculates the superposition value of an object independent of its size. This value can be held constant for the two views of an object and the two objects of a pair, independently of the bag complexity, when combining the bag image and the prohibited item. To ensure that the bag images do not contain any other prohibited items, they were visually inspected by at least two highly experienced aviation security instructors. Clean bag images were assigned to the four categories and the two viewpoints of

the prohibited items such that their image difficulty was balanced across all groups. This was achieved using the false alarm rate as the difficulty indicator for each bag image based on a pilot study with 192 screeners. In the test, each bag appears twice, once containing a prohibited item (threat image) and once not containing a prohibited item (non-threat image). Combined with all prohibited items this adds up to a total of 256 test trials: 4 threat categories (guns, IEDs, knives, other) * 8 (exemplars) * 2 (sets) * 2 (views) * 2 (threat images vs. non-threat images). The task is to inspect visually the test images and to judge whether they are OK (contain no prohibited item) or NOT OK (contain a prohibited item). Usually the images disappear after 15 seconds. In addition to the OK / NOT OK response, screeners have to indicate the perceived difficulty of each image in a 100-point scale (difficulty rating; 1 = easy, 100 = difficult). All responses can be made by clicking buttons on the screen. The X-Ray CAT takes about 30-40 minutes to complete.

1.3.3 Assessing detection performance in a computer-based test

Detection performance of screeners in a computer-based test can be assessed by their judgments of X-ray images. As explained above, not only is the hit rate (i.e., the proportion of correctly detected prohibited items in the threat images) an important value but so is the false alarm rate (i.e., the proportion of non-threat images that were judged as being NOT OK, that is, as containing a prohibited item). This incorporates the definition of detection performance as the ability not only to detect prohibited items but also to discriminate between prohibited items and harmless objects (that is, to recognize harmless objects as harmless). Therefore, in order to evaluate the detection performance of a screener, his or her hit rate in the test has to be considered as well as his or her false alarm rate (Green & Swets, 1966; Hofer & Schwaninger, 2004, 2005; Macmillan & Creelman, 1991). There are different measures of detection performance that set the hit rate against the false alarm rate, for example, *d'* or A'. These measures are explained in more detail below.

1.3.4 Reliability of the X-Ray CAT

As elaborated earlier in this chapter the reliability of a test stands for its repeatability or consistency. The reliability of the X-Ray CAT was measured by computing Cronbach's alpha and Guttman's split-half coefficients. The calculations are based on the results of a study at several airports throughout Europe (see below for the details and further results of the study) including the data of 2265 screeners who completed the X-Ray CAT on behalf of the EU funded VIA project in 2007. The reliability measures were calculated based on correct answers, that is, hits for threat images and correct rejections (CR) for non-threat images (#correct rejections = #non-threat items - #false alarms). The analyses were made separately for threat images and for non-threat images. Table 1.1 shows the reliability coefficients.

Table 1.1. Reliabilities

RELIABILITY ANALYSES			
Reliability Coefficients		Hit	CR
X-Ray CAT	Alpha	.98	.99
	Split-half	.97	.99

As stated above, an acceptable test should reach split-half correlations of at least .75 and Cronbach alpha values of at least .85. Bearing this in mind, the reliability values listed in Table 1.1 show that the X-Ray CAT is very reliable and therefore a useful tool for measuring detection performance of aviation security screeners.

1.3.5 Validity of the X-Ray CAT

Regarding the different types of validity as described above, the face validity and the content validity can be confirmed instantly. In terms of face validity, the X-Ray CAT is valid as it appears to measure what it claims to measure and it reflects the relevant operational conditions. In terms of content validity, the X-Ray CAT is valid as its content is representative of the content of the relevant task. The test includes

prohibited items from different categories, on the basis of the definition in Doc 30 of the European Civil Aviation Conference (ECAC) that have to be detected by the aviation security screeners. Regarding the convergent validity of the CAT, it can be compared to another test that measures the same abilities. An example of such a test that is also widely used at different airports is the Prohibited Items Test (PIT, see Hardmeier, Hofer, & Schwaninger, 2006a, for details). To assess convergent validity, the correlation between the scores on the X-Ray CAT and the scores on the PIT of a sample that conducted both tests is calculated. This precise procedure was applied to a sample of 473 airport security screeners. The result can be seen in Figure 1.4 ($r = .791$).

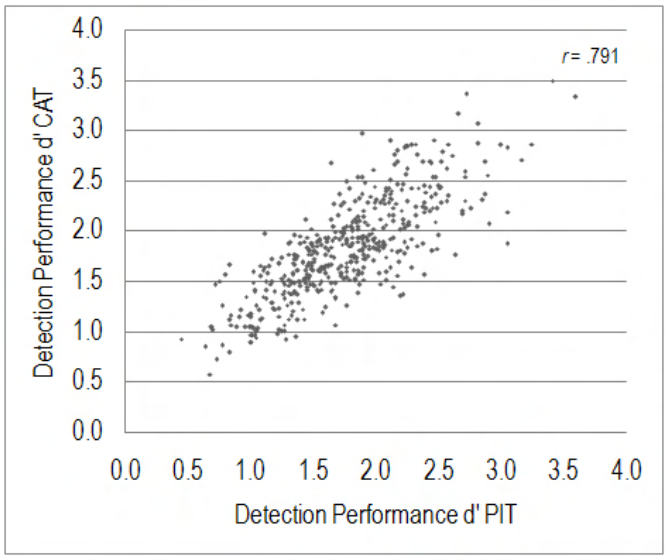


Fig. 1.4. Convergent validity shown as the reliability between the scores of the X-Ray CAT and the PIT. The dots represent individual screeners.

Since correlation coefficients range from 0 (no correlation) to 1 (perfect correlation) (see also above), the convergent validity can be classified as quite high. This means that the X-Ray CAT and the PIT measure the same X-ray image interpretation competency. Other studies have also confirmed the concurrent validity, that is, the ability of a test to discriminate, for example, between trained and untrained screeners (cf. Koller, Hardmeier, Michel, & Schwaninger, 2008). Figure 1.5 shows the results of the study. It can be seen that the detection performance increases for the trained screeners but not for the untrained screeners. This means that the test is able to discriminate between screeners who received training with the computer-

based training system X-Ray Tutor and those who did not receive training with X-Ray Tutor (for details on X-Ray Tutor please refer to Schwaninger, 2004, 2005b). Therefore, the concurrent validity of the X-Ray CAT can be confirmed.

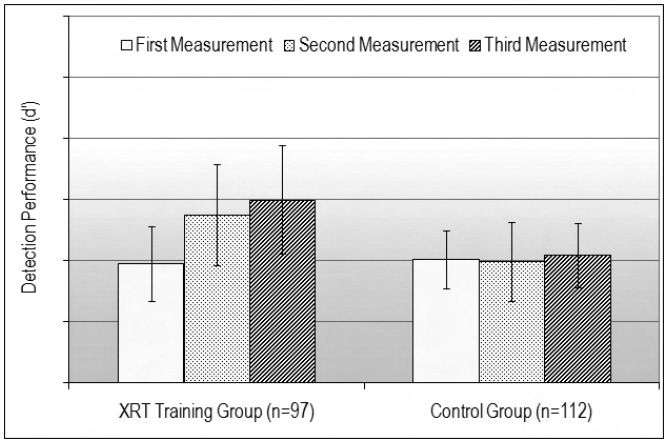


Fig. 1.5. Detection performance *d'* for trained (XRT Training Group) compared to untrained (Control Group) screeners. The concurrent validity appears in the difference of the detection performance between the two groups after one group has trained. Thin bars are standard deviations. Note: No performance measures are indicated for security reasons.

1.3.6 Standardization

The X-Ray CAT was standardized in regard to its development. The revisions of the test were based on data from representative samples (N > 94) of airport security screeners (more details on the revisions can be found in the following subsection). In the study described in section 1.4, involving a large and representative sample of airport security screeners (N = 2265), a mean detection performance A' of 0.8 ($SD = 0.08$) was achieved. There are different approaches to the definition of pass marks. The normative approach defines a pass mark as the threshold where a certain proportion of screeners fails the test (e.g., not more than 10 percent), based on a test measurement. That is, a screener is rated in relation to all other screeners. The criterion-referenced approach sets the pass mark according to a defined criterion. For instance, the results could be compared to the test results obtained in other countries when the test was conducted the first time or by having a group of experts (e.g., using the Angoff method, Angoff, 1971) rate the difficulty of the test items (in this case of the images) and the minimum standard of performance. These approaches can of course be combined. Furthermore, the standard might be adjusted by taking into account the reliability of the test, the confidence intervals, and stan-

dard error of measurement. According to the Measurement Research Associates, the level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation and should not be adjusted to regulate the number or proportion of persons passing the test (*Criterion referenced performance standard setting*, 2004). The pass point should be determined by careful analysis and judgment of acceptable performance. The Angoff method is probably the most basic form of the criterion-based standard setting, due to the relatively simple process of determining the pass points (Khalid & Saeed, 2007). In this method, judges are expected to review each test item and a passing score is computed from an estimate of the probability of a minimally acceptable candidate answering each item correctly. As a first step, the judges discuss and define the characteristics of a minimally acceptable candidate. Then, each judge makes an independent assessment of the probability for each item that this previously defined minimally acceptable candidate will answer the item correctly. To determine the probability of a correct response for each item, that is, the passing score, the judges' assessments of the items are averaged. Then, these probabilities for all items of the test are averaged to obtain the pass point for the test (*Criterion referenced performance standard setting*, 2004). The Angoff method features several advantages: it is easy to implement, understand, and compute (Berk, 1986). However, the Angoff method also has disadvantages. First, it assumes that the judges have a good understanding of the statistical concepts. Second, the panelists may lose sight of the candidates' overall performance on the assessment due to the focus on individual items, as this method uses an item-based procedure (Khalid & Saeed, 2007). Moreover, the continuum of item probabilities tends to result in considerable variability among the judges. Many judges have difficulties defining candidates who are minimally competent (Berk, 1986). In the case of aviation security screeners, judges would have to focus on a person who would be just sufficiently capable of doing this job.

1.3.7 Revision of the test

The development of a scientifically approved test is a complex procedure. Here, the development of the X-Ray CAT is explained in order to provide an example. The first step in a test's development is the definition of what should be measured and how. It was planned that a test should be developed for the purpose of measuring the X-ray image interpretation competency of airport security screeners when they search X-ray images of passenger bags for prohibited objects. In order for the test to be face valid (see above), the nature of the items was obvious. They should be X-ray images of passenger bags where some of these images contain a prohibited item and some do not. Careful thought should be invested into the design of the test. In this case, since it is known that several factors can influence the detection performance of an aviation security screener, the items should be constructed considering these factors. That is, the items should be constructed while controlling for the image-based factors view difficulty, superposition, and bag complexity. Furthermore, the effects that should or could be measured with the test should be considered. Depending on the initial point and the aims, the items can be developed quite differently. The X-Ray CAT is composed of two similar sets and contains prohibited items of different categories, each one in two different viewpoints. The set construction serves the purpose of measuring the transfer effects. Transfer effect means the transfer of knowledge about threat objects that is gained during training to threat objects that were not included in training but are similar to objects that were included. The X-Ray CAT can measure several effects: the effect of viewpoint, threat category, training, and transfer (see above for a more detailed description). After the first version of the test has been constructed, it was administered to a large and representative sample in a pilot study (N = 354 airport security screeners). On the basis of the results of this pilot study, the first revision took place. First of all, a reliability analysis gave information on the quality of the test and each item (item difficulty and item-to-total correlation). Those items with a difficulty below the range of acceptable difficulty had to be revised. The range of acceptable item difficulty depends on the answer type. In this case, an item can be correct or incorrect, that is having a 50 percent

chance probability. The range of acceptable difficulty was defined between 0.6 and 0.9. Furthermore, the items should possess as high an item-to-total correlation as possible. In this case, all items with a negative or very small item-to-total correlation were corrected. In order to measure any effect of threat category on the detection performance, the detection performance of a threat object should only depend on the threat object itself and not on the difficulty of the bag it is placed in. To this end, the difficulty of the bags should be balanced across all categories, across both viewpoints of the test, and also across the two sets. As a measure of difficulty for the bag images, the false alarm rate was consulted (i.e., how many times a bag was judged as containing a threat item although there was none). Then, the bags were assigned to the four categories in such a way that their mean difficulty was not statistically different. The threat objects were built into the new bags if necessary, again considering superposition. Lastly, the items were shifted between the two sets (always incorporating the twin structure) in order to equalize the difficulty of the sets. The revised test was administered to another sample (N = 95 airport security screeners), repeating the revision steps as necessary. After a third (N = 359 airport security screeners) and a fourth (N = 222 airport security screeners) revision, the X-Ray CAT was acceptable in terms of stable reliability, item difficulty, and item-to-total correlation. In summary, the test was revised according to the image difficulty, the item-to-total correlation, and the balancing of the difficulty of the harmless bag images. The aim is to achieve a high reliability with items featuring high item-to-total correlations and acceptable item difficulty. The difficulty of a threat image (a bag containing a prohibited object) should depend only on the object itself and not on the difficulty of the bag. Otherwise, a comparison between the detection performance for the different threat categories could be biased.

1.4 Real World Application of the X-Ray Competency Assessment Test (X-Ray CAT)

X-Ray CAT was used in several studies and in a series of international airports in order to measure the X-ray image interpretation competencies of screening officers.

In this section, the application of X-Ray CAT is presented along with discussions and results obtained by means of the EU-funded VIA Project.

1.4.1 The VIA Project

The VIA Project evolved from the tender call in 2005 of the European Commission's Leonardo da Vinci program on vocational education and training. The project's full title is "Development of a Reference Levels Framework for A'VIA'tion Security Screeners". The aim of the project is to develop appropriate competence and qualification assessment tools and to propose a reference levels framework (RLF) for aviation security screeners at national and cross/sectoral level. To date, 11 airports in six European countries are involved in the project. Most of these airports are going through the same procedure of recurrent tests and training phases. This makes it possible to scientifically investigate the effect of recurrent weekly computer-based training and knowledge transfer and subsequently to develop a reference levels framework based on these outcomes. The tools used for testing in the VIA project are the computer-based training (CBT) program X-Ray Tutor (for details please refer to Schwaninger, 2004, 2005b), and the X-Ray CAT. Subsequently, the results of the computer-based test measurements as part of the VIA project procedure are reported in detail.

1.4.2 VIA computer-based test measurement results

As explained earlier, the X-Ray Competency Assessment Test (X-Ray CAT) contains 256 X-ray images of passenger bags, half of which contain a prohibited item. This leads to four possible outcomes for a trial: a "hit" (a correctly identified threat object), a "miss" (a missed threat object), a "correct rejection" (a harmless bag correctly judged as being OK) and a "false alarm" (an incorrectly reported threat object). In terms of sensitivity, the hit rate alone is not a valid measure to assess X-ray image interpretation competency. It is easy to imagine that a hit rate of 100 percent can be achieved by simply judging every X-ray image as containing a prohibited item. In this case, the entire set of non-threat items is completely neglected

by this measure (the false alarm rate would also be 100 percent). In contrast, Green and Swets (1966) developed a signal detection performance measure *d'* (pronounced 'd prime'), taking into account hit rates as well as false alarm rates. Often, *d'* is referred to as sensitivity, emphasizing the fact that it measures the ability to distinguish between noise (in our case an X-ray image of a bag without a threat) and signal plus noise (in our case an X-ray image containing a prohibited item). *d'* is calculated using the formula $d' = z(H) - z(F)$, where H is the hit rate, F the false alarm rate, and z refers to the z-transformation[6].For the application of *d'*, the data have to fulfill certain criteria (noise and signal plus noise must be normally distributed and have the same variance). If these requirements are not fulfilled, another established, "non-parametric" measure is often used: A' (pronounced 'A prime'). The measure also meets the requirement of setting the hit rate against the false alarm rate in order to achieve a reliable and valid measure of image interpretation competency. A' was the measure of choice for the current analyses because its non-parametric character allows its use independently from the underlying measurements distributions. A' can be calculated as follows, where H represents the hit rate of a test candidate or group and F represents its false alarm rate: $A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)$. If the false alarm rate is larger than the hit rate, the equation must be modified: $A' = 0.5 - [(F - H)(1 + F - H)]/[4F(1 - H)]$. For further information on the mentioned measures and others, please refer to Stanislaw and Todorov (1999), to Green and Swets (1966) for *d'*, and to Pollack and Norman (1964), Grier (1971), or Macmillan and Creelman (1991) for A'. For more details on the application of these formulae in airport security screening please refer to Hofer and Schwaninger (2004). The reported results provide graphical displays of the relative detection performance measures A' at the eight European airports that participated in the present study, as well as another graph showing the effect of the two viewpoints on the different threat categories as explained earlier. In order to provide statistical corroboration of these results, an analysis of variance (ANOVA) on the two within-participants factors, view difficulty and threat category (guns, IEDs, knives, and other items) and the between-participants factor airport is reported as well. As part of the ANOVA,

[6] Z-transformation transforms normal random variables to the standard normal. The standard normal distribution is the normal distribution with a mean of zero and a variance of one.

only the significant interaction effects are reported and considered to be noteworthy in the context.

Detection performance comparison between airports

Figure 1.6 shows the comparison of the detection performance achieved eight European airports that participated in the VIA project. First, the detection performance was calculated for each screener individually. The data were averaged across screeners for each airport. Thin bars represent the standard deviation (a measure of variability) across screeners. Due to its security sensitivity and for data protection reasons, the individual airports' names are not indicated and no numerical data are given here.

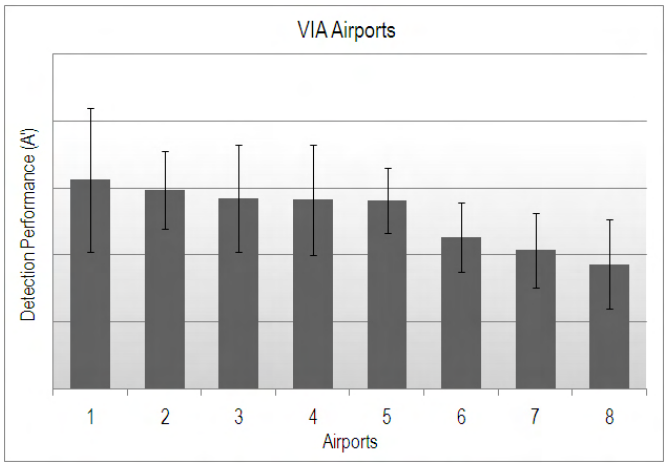


Fig. 1.6. Comparison between eight European airports participating in the VIA project. Thin bars represent standard deviations between screeners.

Although no numerical data is displayed in the graph, we can discern substantial differences between the airports in terms of mean detection performance and standard deviation. As described above, all VIA airports go through a similar procedure of alternation of test phases and training phases. Nevertheless, there are considerable differences between them. There were large differences in the initial positions when the project was started, and the baseline assessment test, which is reported here, was conducted at different times at different airports. The differences can be put down to differences in the amount of training that was accomplished prior to this baseline test as well as on differences in the personnel selection assessment. Some of the re-

ported airports were already coached prior to the VIA project, though with diverse intensity and duration. Taking these differences into account, the reported results correspond fairly well with our expectations based on earlier studies on training effects.

Detection performance comparison between threat categories regarding view difficulty

Figure 1.7 shows again the detection performance measure A', but with a different focus. The data are averaged across airports shown in Figure 1.6, but analyzed by view difficulty within threat categories. There is a striking effect on detection performance deriving from view difficulty. Performance is significantly higher for threat objects depicted in easy views than for threat objects depicted in difficult views (canonical view rotated by 85 degrees).

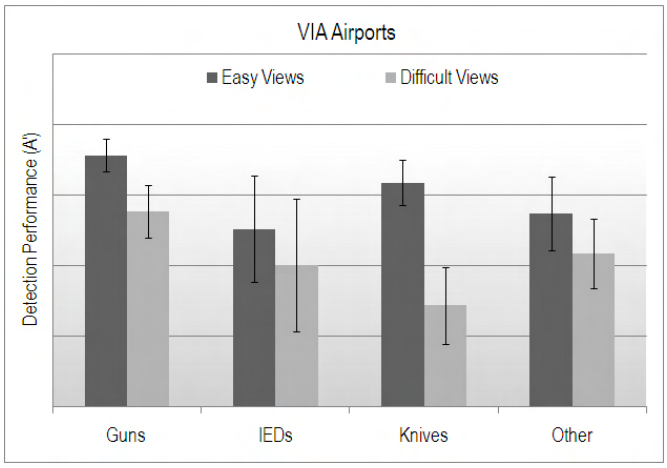


Fig. 1.7. Detection performance A' broken up by category and views (unrotated (easy view) vs. 85° rotated objects (difficult view)). The thin bars represent standard deviations between the eight VIA airports. Pairwise comparisons showed significant viewpoint effects for all four threat categories.

Although this effect can be found in every one of the four threat categories, there are significant differences between them regarding general differences between the mean detection performances and also between the effect sizes of view difficulty that are unequal between threat categories. Knives and IEDs, for example, differ very much in view difficulty effect size but not so much in average detection performance. As can be seen in Figure 1.8, the reason is quite simple: IEDs consist of several parts and not all parts are depicted in easy or in difficult view at same time. Some

parts are always depicted in easy view when others are difficult, and vice versa. Knives have very characteristic shapes. They look consistently longish when seen perpendicular to their cutting edge but very small and thin when seen in parallel to their cutting edge. This interaction effect between threat item category and view difficulty can easily be observed in Figure 1.7, where the difference between easy and difficult views is much larger in knives than in IEDs. Furthermore, based on earlier studies on training effects, it is important to mention here that this pattern shown in Figure 1.7 is also highly dependent on training (Koller et al., 2008) (interaction effects [category * airport and view difficulty * airport]).

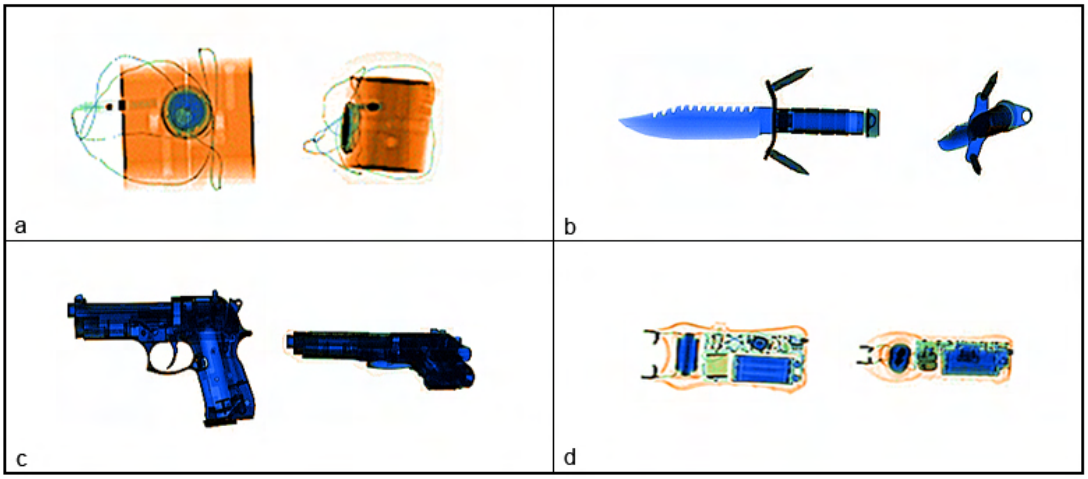


Fig. 1.8. Illustration of how effects of view difficulty differ between the four ECAC threat categories. 8a and b show an IED and a knife each in a frontal view and a rotated view from almost 90 degrees around the vertical axis. 8c and d show a gun and a taser each in a frontal view and a rotated view from almost 90 degrees around the horizontal axis.

Analysis of variance (ANOVA)

The following statistics provide quantitative values of what has been reported graphically. This allows us to compare the effects of the different factors. We applied a three-way ANOVA with the two within-participants factors category and view difficulty, and one between-participants airport factor on the detection performance measure A'. The analysis revealed highly significant main effects

on threat category (guns, IEDs, knives, and other items) with an effect size of $\eta^2 = .131, F(3, 5602.584) = 339.834, MSE = 2.057, p < .001$, on view difficulty (easy view vs. difficult/rotated view) with an effect size of $\eta^2 = .47, F(1, 2257) = 2009.772, MSE = 9.031, p < .001$, and also on the between-participants factor airport with an $\eta^2 = .080, F(1, 2257) = 28.128, MSE = 1.415, p < .001$. The following two-way interactions were also highly significant: threat category * view difficulty: $\eta^2 = .094, F(3, 6542.213) = 233.969, MSE = .931, p < .001$, threat category * airport $\eta^2 = .068, F(3, 5602.584) = 23.411, MSE = .142, p < .001$,and view difficulty * airport $\eta^2 = .159, F(1, 2257) = 60.953, MSE = .274, p < .001$. These results indicate different detection performance for different threat categories and higher detection performance for prohibited items in easy view than for rotated threat items (the effect of viewpoint, Schwaninger et al., 2004). This is consistent with results reported in the view-based object recognition literature (for reviews see, for example, two works by Tarr & Bülthoff, 1995, 1998). The effect sizes were very large according to Cohen's conventions (Cohen, 1988).

1.4.3 Discussion

Although the reported real world application consists of baseline measurement data only, some important features of the X-Ray CAT could be illustrated well. X-Ray CAT allows us to measure and to evaluate the effects of view difficulty and threat objects practically and independently of each other. Furthermore, the X-Ray CAT can be used as a very reliable tool to compare X-ray image interpretation competency of security staff at different airports and other types of infrastructures using X-ray technology for security control procedures.

1.5 Summary and Conclusions

The competency of a screener to detect prohibited items in X-ray images quickly and reliably is important for any airport security system. Computer-based tests, TIP, and to a limited extent covert tests can be used to assess individual competency

in X-ray image interpretation. However, to achieve reliable, valid, and standardized measurements, it is essential that the requirements and principles detailed in this chapter are followed by those who produce, procure, or evaluate competency assessment of the X-ray image interpretation tests of individual screeners. This chapter introduced the competency assessment in airport security screening. In order to achieve a meaningful result the assessment has to meet the criteria of reliability and validity. Furthermore, the assessment has to be standardized to allow the evaluation of screeners' performance in relation to the population norm. Currently, there are three means for assessing X-ray image interpretation competency: covert testing, threat image projection (TIP), and computer-based image tests. Another important feature of maintaining the high level of X-ray baggage screening within aviation security is the initial and recurrent certification of screening personnel. Threat image projection (TIP) as a means to assess X-ray image interpretation competency was illustrated in detail, as well as the conditions that have to be fulfilled in order for TIP to be a reliable and valid instrument. This chapter also focused on the computer-based X-Ray Competency Assessment Test (X-Ray CAT). It features very high reliability scores and its design allows for measuring X-ray image interpretation competency of aviation security screeners with regard to different aspects of their ability and knowledge. The X-Ray CAT is widely used at many different airports throughout the world, for competency assessment and certification purposes as well as in studies assessing the fundamentals of the demands required for the job of the aviation security screener. This chapter continued by showing how a reliable, valid, and standardized test can be used to compare X-ray image interpretation competency across different airports and countries. The results of an EU-funded project (VIA Project) showed remarkable differences in mean detection performance across eight European airports. All these countries currently conduct weekly recurrent computer-based training. Since the X-Ray CAT will be conducted again in the first quarter of 2008, the VIA Project will also provide important insights on the benefits of computer-based training for increasing security and efficiency in X-ray screening.

Acknowledgment

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). This chapter is a summary of Work Package 6: Competency assessment tests. For more information, see www.viaproject.eu.

2

Investigating Training, Transfer, and Viewpoint Effects Resulting from Recurrent CBT of X-ray Image Interpretation

2.1 Introduction

The importance of aviation security has increased dramatically in the last years. As a consequence of the new threat situation, large investments have been made in modern security technology. State of the art X-ray screening equipment offers good image quality, high resolution and many image enhancement functions. However, the decision on whether an X-ray image of a passenger bag contains a suspected item or not, is still being taken by a human operator, that is, an airport security screener. Object shapes that are not similar to ones stored in visual memory are difficult to recognize (e.g., Graf, Schwaninger, Wallraven, & Bülthoff, 2002; Schwaninger, 2004, 2005a). Schwaninger, Hardmeier, and Hofer (2005) have shown that X-ray screener performance depends on knowledge-based and image-based factors. A prerequisite for good X-ray detection performance is knowledge about which objects are prohibited and what they look like in X-ray images. Such knowledge is acquired by computer-based, class-room, and on-the-job training (knowledge-based factors). Image-based factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity due to the number and type of other objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training (Hardmeier, Hofer, & Schwaninger, 2006b). Computer-based training is expected to be a very important determinant of X-ray image interpretation competency, because many threat objects are not known from

everyday experience and because objects look quite different in X-ray images than in reality. This is illustrated in Figure 2.1 with two examples.

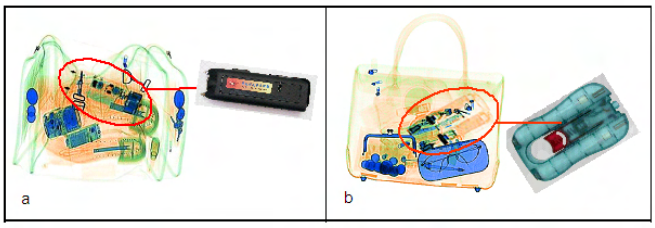


Fig. 2.1. Different types of prohibited items in X-ray images of passenger bags. a Electric shock device, b self defense gas spray "Guardian Angel"

Schwaninger and Hofer (2004) and Schwaninger, Hofer, and Wetter (2007) could show that detection of improvised explosive devices (IEDs) in hold baggage screening (HBS) can be significantly improved if people are trained with an individually adaptive training system such as X-Ray Tutor (XRT). Schwaninger et al. (2005) compared detection performance of novices with the one of aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g., self-defense gas spray, electric shock device, etc.) in X-ray images was found for novices. For experienced aviation security personnel, a much higher recognition performance was observed. McCarley, Kramer, Wickens, Vidoni, and Boot (2004) reported a better performance after training of novices for the detection of knives in X-ray images. When one takes into account the myriad of views that can be produced by a single object, the question arises how the human brain stores and recognizes objects even if they are presented in unusual views. In the object recognition literature, two types of theories can be distinguished: structural description theories and view-based theories. The former assume that objects are stored in visual memory by their component parts and their spatial relationship. An object-centered description of this nature was described by Marr and Nishihara (1978), who proposed that objects are hierarchically decomposed into their parts and spatial relations relative to object-centered coordinates in order to access an object-centered 3D model in visual memory. In the recognition by components (RBC) theory by Biederman (1987), non-accidental properties like vertices, parallel vs. non-parallel lines, straight vs. curved lines etc. (see Lowe, 1985, 1987) are extracted from a line drawing representation of objects to define basic geometrical primitives (geometrical ions, "geons") that

are relatively orientation-invariant. A geon structural description (GSD) in memory is activated by extracting geons from the visual input and matching geon properties and their spatial relationship with the GSD (Hummel & Biederman, 1992). For view-based theories, different approaches have been proposed. Examples are recognition by alignment to a 3D representation (Lowe, 1987), recognition by linear combination of 2D views (Ullman & Basri, 1991), recognition by view interpolation (e.g., using RBF networks) proposed by Poggio and Edelman (1990) and storing of multiple views for each object plus performing transformations (Tarr & Pinker, 1989). What view-based theories have in common is the assumption that objects are not stored in memory as rotation invariant structural descriptions but instead in a format which is viewer-centered. A more detailed discussion of structural description theories vs. view-based theories and more recent hybrid theories is beyond the scope of this paper (for reviews see for example Graf et al., 2002; Hayward, 2003; Kosslyn, 1994; Peissig & Tarr, 2007; Schwaninger, 2005a; Tarr & Bülthoff, 1998). However, it should be pointed out that empirical results seem to be correlated with the required level of recognition (Bülthoff, Edelman, & Tarr, 1995; Tarr, 1995): if the object has to be recognized at 'entry level', behavioral measures are less affected by changes in perspective. However, in the case of subordinate recognition in which fine discrimination is typically required, both response times and accuracy are more sensitive to the specific viewpoint used. Furthermore, differences in the task a subject has to perform (Lawson, 1999) and the specific paradigm that is used (Verfaillie, 1992) can influence which level of representation is tapped (see also Logothetis & Sheinberg, 1996). The first aim of this study is to investigate how well airport security screeners can detect guns, knives, IEDs, and other prohibited items in X-ray images of passenger bags. The second aim is to examine whether screener detection performance can be increased by conducting recurrent CBT. To this end, screeners conducted weekly recurrent CBT (about 20 min per week). Detection performance was tested with the X-Ray Competency Assessment Test (X-Ray CAT) by Koller and Schwaninger (2006). This test measures how well people detect threat items in X-ray images of passenger bags. It was conducted at the beginning and then after three and six months of training. In addition to training effects, the X-Ray CAT

allows measuring transfer effects, that is, to what extent visual knowledge that was gained through CBT can be transferred to other threat items (see below). In the X-Ray CAT all prohibited items are depicted from a canonical (easy recognizable) perspective (Palmer et al., 1981) and unusual perspective which allows investigating viewpoint effects. The study was conducted at two mid-size European airports. In Airport 1 (Experiment 1) one group of screeners used adaptive CBT (XRT) whereas the other group of screeners (control group) used a conventional (not adaptive) CBT. In Airport 2 (Experiment 2) the same experimental design was used except for the fact that the control group used another conventional CBT system. This allows investigating whether a training effect is dependent on the type of the CBT system used.

2.2 Experiment 1

2.2.1 Method

Participants

A total of 209 airport security screeners of a mid-size European airport participated in Experiment 1 and conducted the X-Ray CAT three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 97 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 112 screeners who used a conventional (not adaptive) CBT. According to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.26 minutes (SD = 3.65 min) per week.

Materials and Procedure

X-Ray Competency Assessment Test (X-Ray CAT)

The X-Ray CAT consists of 256 trials based on 128 different color X-ray images of passenger bags. Each of the bag images is used once containing a prohibited item (threat image) and once without any threat object (non threat image). Figure 2.2 displays examples of the stimuli. Note that in the test the images are displayed in color.

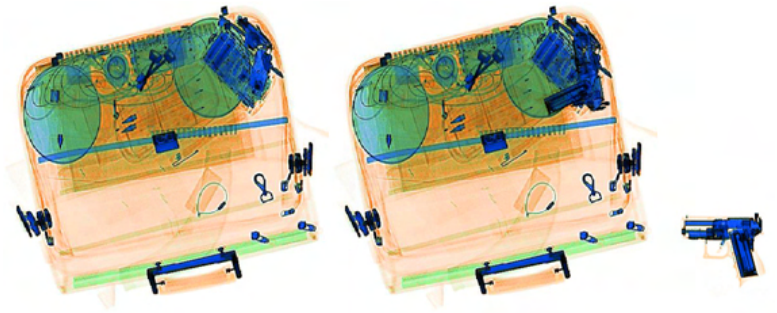


Fig. 2.2. Example images from the X-Ray CAT. Left: harmless bag (non threat image), right: same bag with a prohibited item at the top right corner (threat image). The prohibited item (gun) is shown also separately at the bottom right.

Prohibited objects can be assigned to four categories as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, IEDs, knives, and other prohibited items (e.g., self-defense gas spray, chemicals, grenades, etc.). The threat objects were selected and prepared in collaboration with airport security experts to be representative and realistic. For each threat category 16 exemplars are used (eight pairs). Each pair consists of two prohibited items that are similar in shape (see Figure 2.3). These were distributed randomly into two sets, set A and set B. Prohibited items of set A (without bags) are contained in the XRT CBS 2.0 SE training whereas the items of set B are not. This allows testing for transfer effects.

Every item is depicted from two different viewpoints. The easy viewpoint refers to the canonical (i.e., easy recognizable) perspective (Palmer et al., 1981). The difficult viewpoint shows the threat item with an 85 degree horizontal rotation or an 85 degree vertical rotation relative to the canonical view (see Figure 2.3 for examples). In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally. Set A and B are equalized for the rotations of

Fig. 2.3. Example of two X-ray images of similar looking threat objects used in the test. Left: a gun of set A. Right: corresponding gun of set B.

the prohibited objects. Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula:

$$SP = \frac{\sqrt{\sum [I_{SN}(x,y) - I_N(x,y)]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); ObjectSize: Number of pixels of the prohibited item where R, G, and B are < 253

Using this equation (note the division by object size), the superposition value is independent of the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independently of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories[1]. The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners of another airport. The X-Ray CAT takes about 30-40 minutes to complete. Each image is shown for a maximum of 15 seconds on the screen. Screeners have to judge whether the bag

[1] The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints can not be due to differences in the difficulty of the bag images.

is OK (contains no prohibited item) or NOT OK (contains a prohibited item). Additionally, screeners have to indicate the perceived difficulty of each image on a 100 point scale (difficulty rating)[2].The X-Ray CAT is built into the XRT training system (see below). The interface of the X-Ray CAT is the same as in XRT except there is no feedback and screeners do not have to click on the image to identify the threat object.

X-Ray Tutor (XRT) Training System

X-Ray Tutor (XRT) is an individually adaptive training system for aviation security screeners. It contains a large image library with hundreds of different threat objects depicted in up to 72 views, more than 6000 bag images and many millions of possible threat object to bag combinations (see Schwaninger, 2004, for details).

The individually adaptive training algorithm of XRT starts with showing threat objects depicted from easy viewpoints with little superposition by other objects and in bags of low complexity. On the basis of each individual screener's learning progress, threat objects are shown in more difficult views, in more complex bags and with more superposition. These parameters are adapted automatically by a scientifically validated algorithm for each screener and threat object, which uses automatic image processing algorithms as explained in Schwaninger, Michel, and Bolfing (2007). XRT first presents screeners prohibited objects in easy (canonical) views. The individually adaptive training algorithm determines for each screener which views are difficult to recognize and adapts the training so that the trainee becomes able to detect threat items reliably even if prohibited objects are substantially rotated away from the easiest view. During the next difficulty levels, first superposition and then bag complexity is increased so that the trainee becomes able to detect threat items reliably even if they are superimposed on other objects or if the complexity of a bag is very high (for more information on XRT see Schwaninger, 2003a, 2004, 2005b). During a training session, each image is displayed for 15 seconds on the screen. Within this time screeners can use image enhancement functions which are also available when working with the X-ray machine (e.g., grayscale, negative image, edge enhancement,

[2] The difficulty ratings were not analyzed in this study.

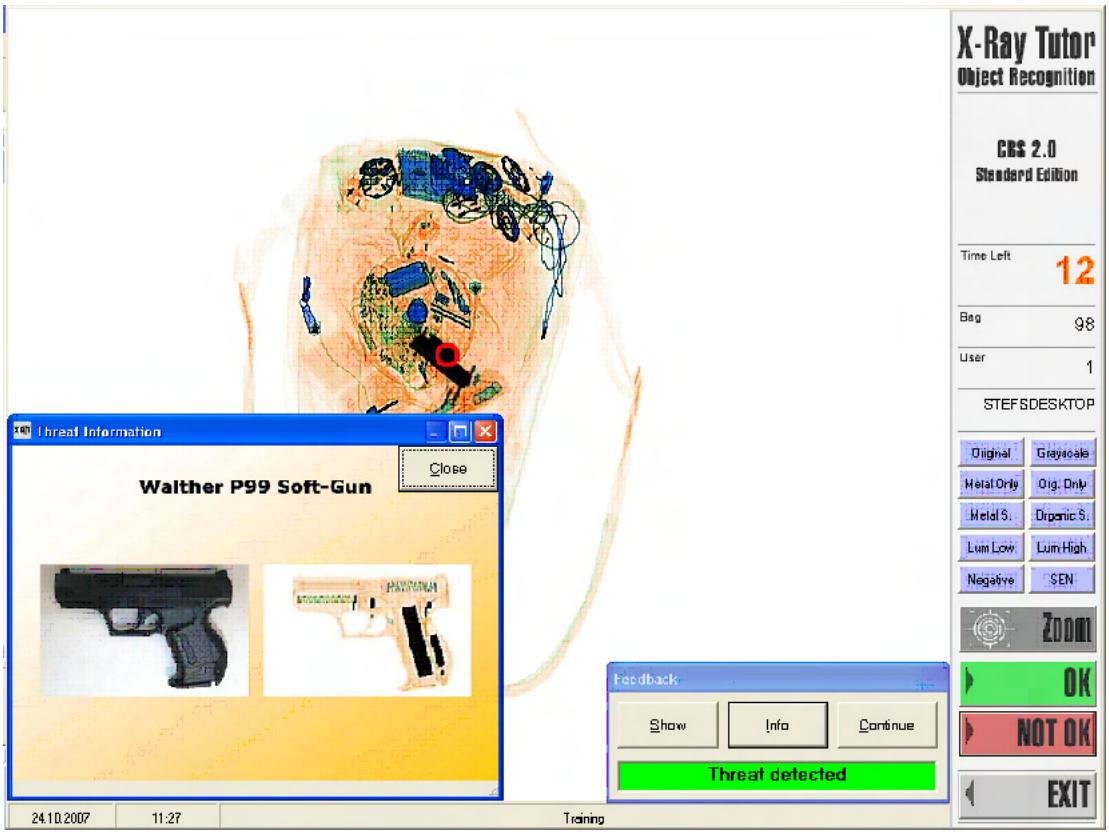


Fig. 2.4. Screenshot of the XRT CBS 2.0 training system during training. At the bottom right a feedback is provided after each response. If a bag contains a prohibited item, an information window can be displayed (see bottom left of the screen).

etc.). If the image contains a prohibited item, screeners have to click on it and then click on the NOT OK button. If the bag is harmless, they have to click on the OK button. After providing a confidence rating using a slider control, feedback is shown to inform the trainee whether the image has been judged correctly or not (see Figure 2.4). If the bag contains a threat item, it is highlighted by flickering and the trainee has the possibility to display information about the threat item (see bottom left of Figure 2.4). By clicking on the 'continue' button the next image is shown. As a default setting, one training sessions takes 20 minutes. During this time screeners see between 150 and 300 images.

Procedure

As explained above, two groups of screeners participated in Experiment 1. The XRT training group conducted weekly recurrent CBT using XRT CBS 2.0 Standard Edition. The control group used a conventional (not adaptive) CBT. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article. However, it can be mentioned that this CBT is also widely used at many airports worldwide. It has a much smaller threat image library than XRT; threat objects are not displayed in many different views; threat objects are not matched with different bags on the fly; and there is no individually adaptive training algorithm. The XRT training group and the control group took the X-Ray CAT prior to the beginning of training, after three months, and after six months of weekly CBT. This allowed testing the effectiveness of both CBT systems for increasing X-ray image interpretation competency of airport security screeners. As explained above, half of the prohibited items in the X-Ray CAT are also contained in the XRT training system (although presented in different bags). The other half of the prohibited items of the X-Ray CAT is not part of the XRT training library. This allows testing for transfer effects, that is, testing whether training with the detection of certain prohibited items helps increasing the detection of other prohibited items. Finally, as specified above in the section on the X-Ray CAT, all prohibited items are depicted in easy and difficult views which allow testing effects of viewpoint on screener detection performance.

2.2.2 Results and Discussion

Detection performance was calculated using the signal detection measure *d'* (Green & Swets, 1966), which takes into account the hit rate (correctly judged threat images as being NOT OK) and the false alarm rate (wrongly judged harmless bags as being NOT OK). *d'* is calculated using the following formula: $d' = z(H) - z(FA)$. Whereas H is the hit rate, FA the false alarm rate and z refers to the z-transformation. Performance values are not reported for security reasons. However, effect sizes are reported for all relevant analyses and interpreted on the basis suggested by Cohen

(1988), see Table 2.1. For t-tests, d between 0.20 and 0.49 represents small effect size; d between 0.50 and 0.79 represents medium effect size; $d \geq 0.80$ represent large effect size. For analysis of variance (ANOVA) statistics, η^2 between 0.01 and 0.05 represents small effect size; η^2 between 0.06 and 0.13 represents medium effect size; $\eta^2 \geq 0.14$ represent large effect size.

Table 2.1. Classification of effect sizes from Cohen (1988)

Effect size	d	η^2
small	0.20-0.49	0.01-0.05
medium	0.50-0.79	0.06-0.13
large	≥ 0.80	≥ 0.14

Figure 2.5 shows the detection performance of the first, second, and third measurement for both screener groups. As can be seen in the figure, there was a large improvement as a result of training in the XRT training group while there was no improvement in the control group. These results were confirmed by an

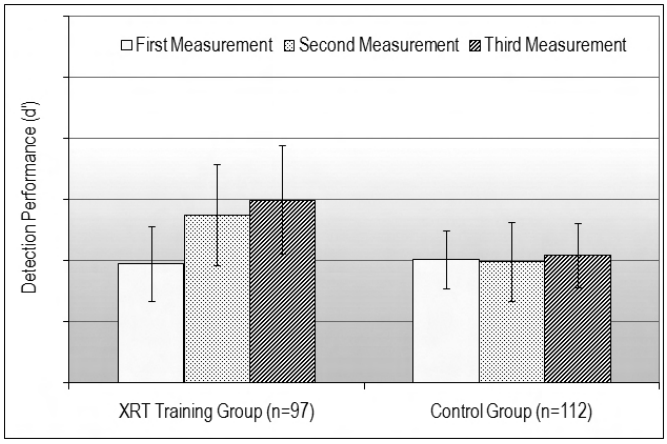


Fig. 2.5. Detection performance with standard deviations for the XRT training group (left) vs. the control group (right) comparing first, second, and third measurement.

ANOVA for repeated measures using *d'* scores with the within-participants factor measurement (first, second, and third) and the between-participants factor group (XRT training group and control group). There were large main effects of measurement, $\eta^2 = .28, F(2, 414) = 81.04, p < .001$, and group, $\eta^2 = .19, F(1, 207) = 47.62, p < .001$. There was also a large interaction of measurement and group, $\eta^2 = .25, F(2, 414) = 68.67, p < .001$, which is consistent with Figure 2.5 show-

ing large performance increases as a result of training only for the XRT training group but not for the control group.

Separate pairwise t-tests of detection performance d' revealed no significant difference at the baseline measurement between the two groups $t(177) = -0.91, p = .363, d = 0.13$, but already a significant difference in the second measurement, that is, after three months of training, $t(207) = 7.52, p < .001, d = 1.04$. Additional paired-samples t-tests revealed significant differences for the XRT training group between all three test measurements but no significant differences for the control group (see Table 2.2).

	$t(96)$	p	d
XRT Training Group (t1 - t2)	-9.80	$< .001$	1.12
XRT Training Group (t2 - t3)	-3.95	$< .001$	0.28
	$t(111)$	p	d
Control Group (t1 - t2)	0.54	$= .59$	0.05
Control Group (t2 - t3)	-1.89	$= .06$	0.17

Table 2.2. Results of the t-tests comparing the detection performance of first (t1), second (t2), and third (t3) measurement

Figure 2.6 shows the detection performance of both screener groups broken down by prohibited item category and the three test measurements. A repeated-measures ANOVA with the within-participants factors measurement (first, second, and third) and threat category (guns, IEDs, knives, and other), and the between-participants factor group (XRT training vs. control) revealed the significant main effects and significant interactions given in Table 2.3a. In addition to the effects that were already found in the previous ANOVA, also the factor threat (or prohibited item) category was significant.

Figure 2.6 shows that, at the first test measurement, guns were detected best, followed by knives, other prohibited items and then by IEDs. There was a highly significant interaction between threat category and measurement. As can be seen in Figure 2.6, detection of IEDs was initially much lower than gun detection. After six months of training, screeners of the XRT training group could detect IEDs

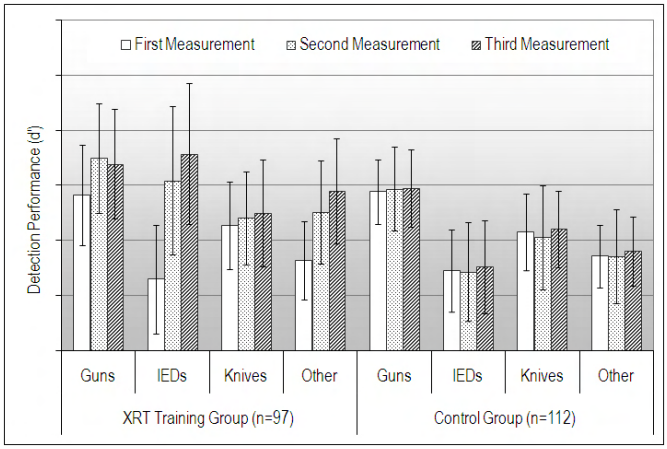


Fig. 2.6. Detection performance with standard deviations for the XRT training group vs. the control group broken down by prohibited item category and test measurement.

even slightly better than guns. This result implies that IED detection is not difficult per se but rather a matter of the right training. Note that in this study all IEDs contained a detonator, wires, explosive, a triggering device, and a power source. Thus our conclusions are only applicable to the detection of multi-component IEDs. Large performance increases were also found for other prohibited items in this group, while for knives, only a small improvement as a result of training was found. Note that after six months of training, performance for knife detection is lower than the one for any other threat category in the XRT training group; although at baseline measurement it was higher than the performance for IED detection or other threat objects. The interaction between threat category, group and measurement is also worth mentioning. As can be seen in Figure 2.6 this results from the fact that there was no training effect for the control group. Their detection performance remains at about the same level for each threat category even after six months of training with the conventional (not adaptive) CBT system. Separate pairwise t-tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately (Table 2.4). The XRT training group showed a significant increase of the detection performance at the second measurement for the categories guns, IEDs, and other threat objects. For knives, a significant difference could be found only in the third measurement. The comparison of the effect size d between the t-tests of the four threat categories confirms the earlier mentioned conclusion that the training effect was particularly big for IEDs and rather small for knives. Detection performance of the control group did not differ

significantly between the measurements, confirming that the conventional CBT did not result in an increase of threat detection performance.

Table 2.3: Results of the ANOVAs in Experiment 1

	Factor	**df**	**F**	η^2	***p***
a)	Measurement (M)	2, 414	83.96	.29	$< .001$
	Threat Category (T)	3, 621	240.03	.54	$< .001$
	Group (G)	1, 207	56.20	.21	$< .001$
	M x G	2, 414	70.49	.25	$< .001$
	T x G	3, 621	45.05	.18	$< .001$
	M x T	6, 1242	43.20	.17	$< .001$
	M x T x G	6, 1242	40.65	.16	$< .001$
b)	Measurement (M)	2, 414	80.55	.28	$< .001$
	Set (S)	1, 207	4.18	.02	$< .05$
	Group (G)	1, 207	49.40	.19	$< .001$
	M x G	2, 414	67.99	.25	$< .001$
	M x S	2, 414	8.80	.04	$< .001$
	S x G	1, 207	51.32	.20	$< .001$
	M x S x G	2, 414	11.54	.05	$< .001$
c)	Measurement (M)	2, 414	87.69	.30	$< .001$
	Set (S)	1, 207	2.37	.01	$= .13$
	Threat Category (T)	3, 621	236.79	.53	$< .001$
	Group (G)	1, 207	63.57	.24	$< .001$
	M x G	2,414	71.16	.26	$< .001$
	M x T	6, 1242	44.35	.18	$< .001$
	M x S	2, 414	10.93	.05	$< .001$
	S x G	1, 207	52.25	.20	$< .001$
	S x T	3, 621	74.00	.26	$< .001$
	T x G	3, 621	47.39	.19	$< .001$
	M x T x G	6, 1242	41.04	.17	$< .001$
	M x S x G	2, 414	10.74	.05	$< .001$
	M x S x T	6, 1242	3.84	.02	$< .01$
	S x T x G	3, 621	4.78	.02	$< .01$
	M x S x T x G	6, 1242	2.99	.01	$< .01$

Continued on Next Page...

Table 2.3: Results of the ANOVAs in Experiment 1

	Factor	df	F	η^2	*p*
d)	Measurement (M)	2,414	84.10	.29	$< .001$
	View (V)	1, 207	1768.63	.90	$< .001$
	Threat Category (T)	3, 621	258.62	.56	$< .001$
	Group (G)	1, 207	61.91	.23	$< .001$
	M x G	2, 414	65.80	.24	$< .001$
	M x T	6, 1242	41.33	.17	$< .001$
	M x V	2, 414	2.05	.01	$= .13$
	V x G	1, 207	3.27	.02	$= .07$
	V x T	3, 621	425.64	.67	$< .001$
	T x G	3, 621	40.86	.17	$< .001$
	M x T x G	6, 1242	40.25	.16	$< .001$
	M x V x G	2, 414	2.23	.01	$< .05$
	M x V x T	6, 1242	6.58	.03	$< .001$
	V x T x G	3, 621	3.08	.02	$< .05$
	M x V x T x G	6, 1242	2.68	.01	$< .05$

The results of the analyses considering the two prohibited item sets of the X-Ray CAT, set A and set B, are shown in Figures 2.7 and 2.8. As explained above, set A items are X-Ray CAT images which contain prohibited items which are part of the XRT image library. Set B items are X-Ray CAT images which contain prohibited items that are not part of the XRT image library. By comparing training effects for set A and set B transfer effects can be investigated, that is, whether training with XRT does not only improve detection of prohibited items that are part of the XRT image library (set A) but also the detection of other prohibited items that are visually similar (set B). Figure 2.7 shows the detection performance for both screener groups broken down by test set for all three measurements. It shows a clear increase in detection performance for the XRT training group, especially at

Table 2.4. Results of the t-tests comparing the detection performance of the four categories between the first (t1), second (t2), and third (t3) measurement.

XRT training group	$t(96)$	**df**	p	d
Guns t1 - t2	- 5.96	96	$< .001$	0.70
IEDs t1 - t2	- 13.03	96	$< .001$	1.53
Knives t1 - t2	- 1.51	96	$= .13$	0.17
Other t1 - t2	- 8.47	96	$< .001$	1.07
Guns t1 - t3	- 4.69	96	$< .001$	0.60
IEDs t1 - t3	- 15.88	96	$< .001$	2.00
Knives t1 - t3	- 2.27	96	$< .05$	0.26
Other t1 - t3	- 12.56	96	$< .001$	1.51
Control group	$t(111)$	**df**	p	d
Guns t1 - t2	- 0.40	111	$= .69$	0.05
IEDs t1 - t2	0.03	111	$= .98$	0.00
Knives t1 - t2	0.83	111	$= .41$	0.09
Other t1 - t2	-0.17	111	$= .87$	0.02
Guns t1 - t3	-0.92	111	$= .36$	0.10
IEDs t1 - t3	-1.05	111	$= .30$	0.08
Knives t1 - t3	-0.73	111	$= .47$	0.08
Other t1 - t3	-1.39	111	$= .17$	0.15

the second measurement, after the first three months of training. For the control group, as in the previous analysis, no training effect is evident.

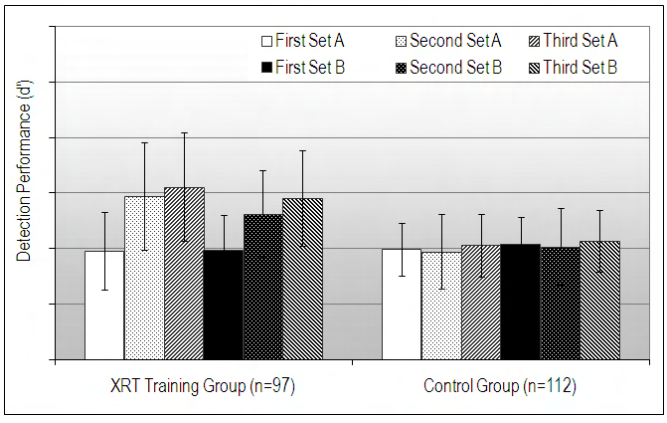


Fig. 2.7. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement for set A and set B separately.

The results of the repeated measures ANOVA with the within-participants factors measurement (first, second, and third) and set (A vs. B) and the between-

participants factor group (XRT training group vs. control group) can be seen in Table 2.3b. There was a significant effect of set in this analysis, which would imply a different detection performance for set A vs. set B. However, the effect is very small, as the effect size of $\eta^2 = 0.2$ clearly shows, which makes the difference quasi negligible. This is also supported by the small effect size for the interaction between set and measurement, $\eta^2 = 0.4$. Pairwise t-tests showed a significant increase in detection performance at the second measurement for both sets for the XRT training group, set A, $t(96) = -10.27, p < .001, d = 1.19$, set B, $t(96) = -7.68, p < .001, d = 0.92$. These results indicate a large transfer effect, that is, visual knowledge regarding the visual appearance of the prohibited objects of the XRT image library helped screeners to detect similar looking, but untrained objects in the X-Ray CAT (set B). Consistent with previous analyses, there was no training effect for the control group, neither for set A, $t(111) = .76, p = .45, d = 0.08$, nor for set B, $t(111) = -0.28, p = .78, d = 0.03$. Pairwise t-tests comparing both sets within one group at the first measurement revealed a significant difference of the two sets only for the control group $t(111) = -2.82, p < .01, d = 0.17$ but not for the XRT training group, $t(96) = -0.42, p = .68, d = 0.03$. However, note that an effect size of $d = 0.17$ is very small which supports the assumption that the two sets are in fact very similar in their difficulty level. Figure 2.8 includes also the threat category in the analysis. The increase in detection performance for the XRT training group can also be seen in the different threat categories.

Pairwise t-tests between the first and second measurement confirmed a significant ($p < .001$, all $d > 0.62$) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A: $p = .12, d = 0.19$, set B: $p = .32, d = 0.12$). In Figure 2.8, detection performance in Set A for guns shows a decrease between the second and third measurement. However, this difference was not significant ($p = .13, d = 0.17$). For the control group, detection performance between the first and third measurement was compared in order to maximize the chances for finding a significant training effect. Even here, for all categories in each set, the detection between the first and third measurement did not differ significantly (all $p > .12, d < 0.18$). The extended ANOVA with the additional

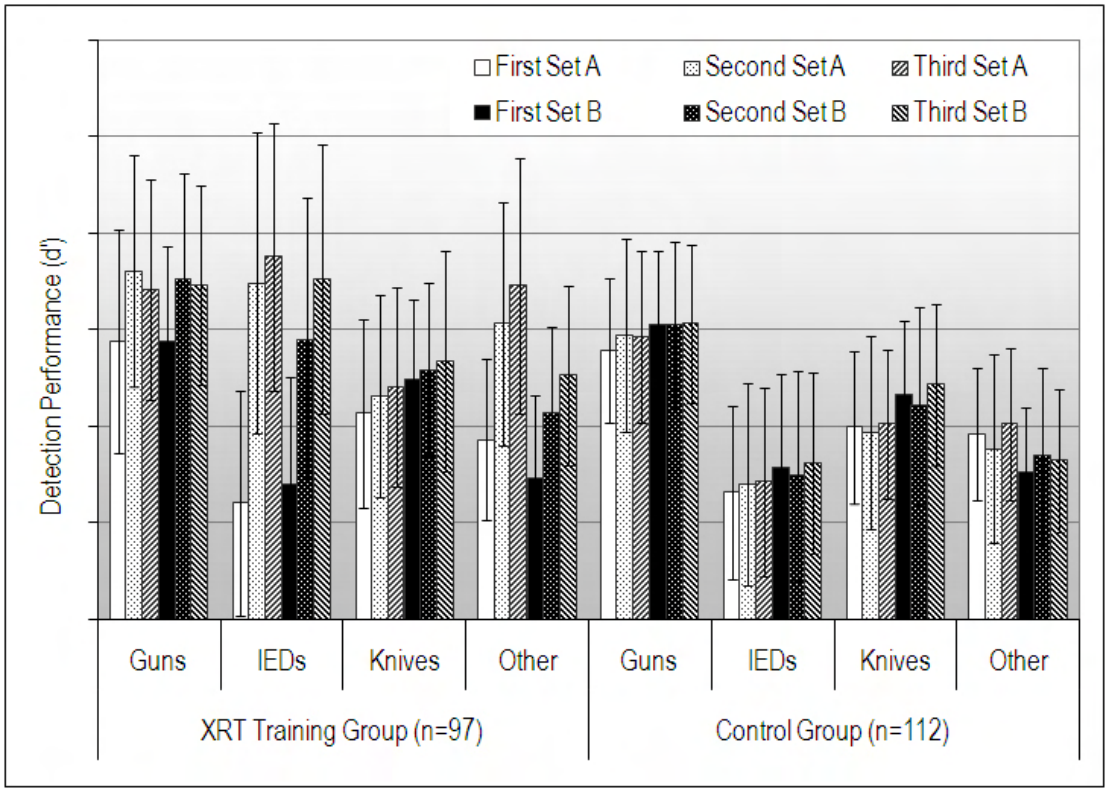


Fig. 2.8. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement for set A and set B and each threat category separately.

within-participants factor threat category revealed the main effects and interactions as specified in Table 2.3c. The main effect of set was not significant but there were significant interactions with set (see Table 2.3c). However, as can be seen in Figure 2.8, these interactions are rather small, which implies large transfer effects. Figure 2.9 shows the results of the viewpoint analysis. An ANOVA was conducted on *d'* scores with the within-participants factors measurement, threat category, and viewpoint and the between-participants factor group. It showed significant main effects of measurement, category, viewpoint, and group. For details and interactions see Table 2.3d. The large main effect of viewpoint indicates a higher detection performance for objects in easy (canonical) viewpoint compared to objects presented in a difficult (rotated) view Figure (cf. Figure 2.9).

However, no significant interaction between viewpoint and training could be found. This would suggest that the viewpoint effect is unaffected by the training and could

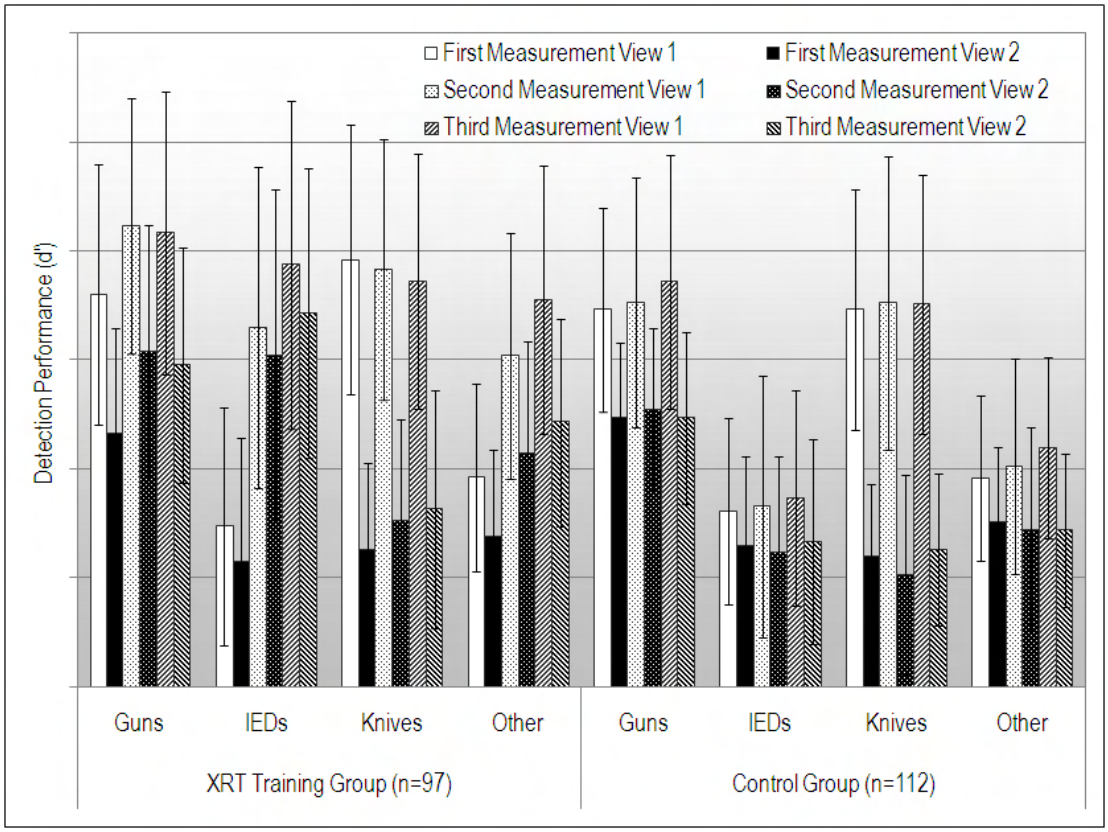


Fig. 2.9. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement for both views and each threat category separately.

not be decreased. Pairwise t-tests showed a significant increase in detection performance at the second measurement for both views in all categories for the XRT training group with the exception of knives in the easy view ($p = .53, d = 0.07$). All other comparisons were significant $p < .05, d > 0.31$). For the control group no significant increase in detection performance could be found (all $p > .10, d < .0.19$), see Table 2.5 for details. Training with XRT has an effect not only on the objects in the easy view but also on those in the difficult view. The screeners could make the association between the rotated object they detected during training and the canonical view of the object which is displayed in the object information in XRT.

In summary, a large and significant training effect was found for the group who trained with XRT for three and six months compared to a control group who used another CBT for the same time. A significant training effect has been observed for all four threat categories (guns, knives, IEDs, and other), whereas the extent of the

Table 2.5. Results of the t-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	$t(96)$	p	d
Guns: V1t1 - V1t2	-4.21	$< .01$	0.53
IEDs: V1t1 - V1t2	-12.25	$< .001$	1.42
Knives: V1t1 - V1t2	0.64	$= .53$	0.07
Other: V1t1 - V1t2	-8.95	$< .001$	1.12
Guns: V2t1 - V2t2	-6.03	$< .001$	0.70
IEDs: V2t1 - V2t2	-11.45	$< .001$	1.43
Knives: V2t1 - V2t2	-2.53	$< .05$	0.31
Other: V2t1 - V2t2	-6.17	$< .001$	0.84
Control group	$t(111)$	p	d
Guns: V1t1 - V1t2	-0.21	$= .84$	0.02
IEDs: V1t1 - V1t2	-0.76	$= .45$	0.08
Knives: V1t1 - V1t2	-0.66	$= .51$	0.07
Other: V1t1 - V1t2	-1.26	$= .21$	0.13
Guns: V2t1 - V2t2	-0.67	$= .50$	0.09
IEDs: V2t1 - V2t2	0.71	$= .48$	0.07
Knives: V2t1 - V2t2	1.65	$= .10$	0.19
Other: V2t1 - V2t2	0.64	$= .53$	0.07

effect varied between categories. A large transfer of the acquired knowledge about the visual appearance of objects used in training (set A) to similar looking objects not used in the training (set B) was found for the XRT training group but not for the control group. This means that training with XRT helped screeners to detect other prohibited items which were not part of the XRT training. Substantial effects of viewpoint could be observed, that is, unusual views of prohibited objects were much harder to detect than canonical views.

2.3 Experiment 2

The main aim of Experiment 2 was to replicate the results of Experiment 1 at another European airport. In addition, another conventional CBT was used for the

control group. Thus it could be investigated whether conventional CBTs differ from each other regarding training effectiveness compared to XRT.

2.3.1 Method

Participants

A total of 163 airport security screeners of another mid-size European airport participated in Experiment 2. All screeners conducted the X-Ray CAT three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 84 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 79 screeners and they used another conventional CBT than the control group of Experiment 1. As in Experiment 1, according to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.92 minutes ($SD = 2.87$) per week.

Materials and Procedure

Materials and procedure in Experiment 2 were the same as in Experiment 1. Again, all screeners took the X-Ray CAT at the beginning and after three and six months of CBT. The only difference was the CBT for the control group, which was not used in Experiment 1. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article for Experiment 2, either. However, it can be mentioned that also this CBT is widely used at many airports worldwide. As the conventional CBT used in Experiment 1, this CBT has a much smaller threat image library than XRT; threat objects are not displayed in many different views; threat objects are not matched with different bags automatically on the fly; and there is no individually adaptive training algorithm.

2.3.2 Results and Discussion

This section is structured the same way as in Experiment 1. Figure 2.10 shows the detection performance *d'* for both groups and all three test measurements. As in Experiment 1, individual *d'* scores were subjected to a repeated measures ANOVA with the within-participants factor measurement (first, second, and third) and the between-participants factor group (XRT training group and control group). Again, there were large main effects of measurement $\eta^2 = .50, F(2, 322) = 163.52, p < .001$, group, $\eta^2 = .26, F(1, 161) = 56.34, p < .001$, and a significant interaction of measurement and group $\eta^2 = .33, F(2, 322) = 78.40, p < .001$.

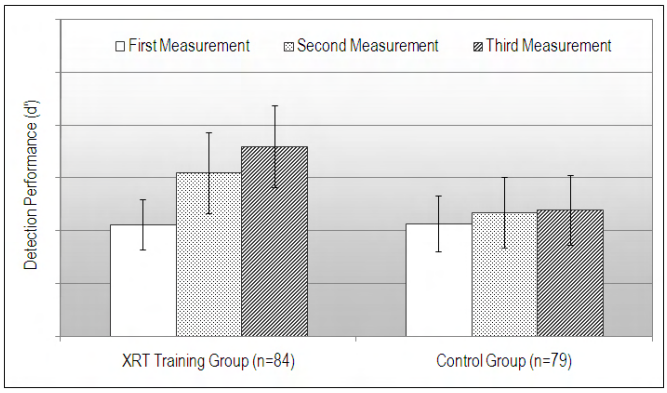


Fig. 2.10. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement.

The large interaction is consistent with Figure 2.10 showing a much larger performance increase as a result of training for the XRT training group when compared to the control group. This was confirmed by independent samples *t*-tests. There was no significant difference between both groups for the first measurement $t(161) = -.22, p = .83, d = 0.03$, but a highly significant difference already in the second measurement $t(161) = 6.66, p < .001, d = 1.05$ after three months of training. As in Experiment 1, additional paired-samples *t*-tests revealed significant differences for the XRT training group between all measurements. In contrast to Experiment 1, there were also significant differences for the control group between the first and second measurement, although not between the second and third measurement (see Table 2.6). Thus, the conventional CBT used in Experiment 2 did also result in increased detection performance although substantially less than XRT.

Table 2.6. Results of the t-tests comparing the detection performance of first (t1), second (t2), and third (t3) measurement

	$t(83)$	p	d
XRT Training Group (t1 - t2)	-12.21	$< .001$	1.57
XRT Training Group (t2 - t3)	-7.07	$< .001$	0.65
	$t(78)$	p	d
Control Group (t1 - t2)	-3.67	$< .001$	0.36
Control Group (t2 - t3)	-0.91	$= .37$	0.07

Figure 2.11 shows the detection performance of both screener groups broken down by prohibited item category and the three test measurements. Again, a clear effect of training on the detection performance can be seen for the XRT training group with the largest increase after the first three months of training. However, also the control group shows a slight increase in detection performance at least for the second measurement.

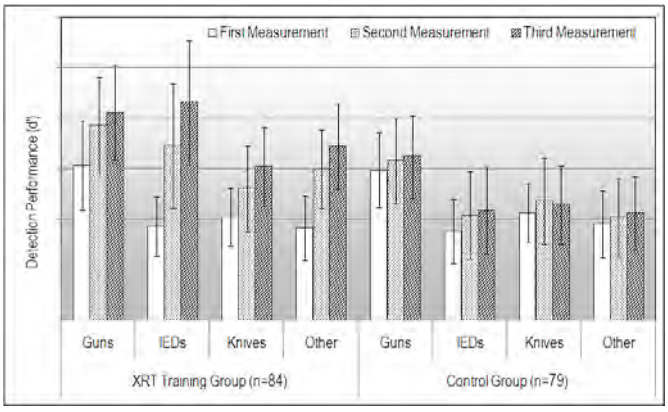


Fig. 2.11. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for each threat category separately.

The analysis of variance (ANOVA) with threat category as additional within-participants factor showed significant main effects and significant interactions (for details see Table 2.7a). The results are comparable to those in Experiment 1. Most importantly, detection of guns was best initially, while detection of IEDs was much lower. After six months of recurrent adaptive CBT, screeners of the XRT training group could detect IEDs even slightly better than guns. This nice replication of the results obtained in Experiment 1 clearly shows that IED detection is not difficult per se but only a matter of the right training. As mentioned above, all IEDs used in

this study contained a detonator, wires, explosive, a triggering device, and a power source. Thus our conclusions are only applicable to the detection of multi-component IEDs.

Table 2.7: Results of the ANOVAs in Experiment 2

	Factor	**df**	**F**	η^2	***p***
a)	Measurement (M)	2, 322	160.78	.50	$< .001$
	Threat Category (T)	3, 483	234.85	.59	$< .001$
	Group (G)	1, 161	64.98	.29	$< .001$
	M x G	2, 322	78.54	.33	$< .001$
	T x G	3, 483	37.63	.19	$< .001$
	M x T	6, 966	26.24	.14	$< .001$
	M x T x G	6, 966	16.67	.09	$< .001$
b)	Measurement (M)	2, 322	156.12	.49	$< .001$
	Set (S)	1, 161	58.45	.27	$< .001$
	Group (G)	1, 161	56.03	.26	$< .001$
	M x G	2,322	82.16	.34	$< .001$
	M x S	2, 322	8.88	.05	$< .001$
	S x G	1, 161	31.37	.16	$< .001$
	M x S x G	2, 322	15.52	.09	$< .001$
c)	Measurement (M)	2, 322	162.28	.50	$< .001$
	Set (S)	1, 161	41.88	.21	$< .001$
	Threat Category (T)	3, 483	231.83	.59	$< .001$
	Group (G)	1, 161	71.93	.31	$< .001$
	M x G	2, 322	84.18	.34	$< .001$
	M x T	6, 966	27.50	.15	$< .001$
	M x S	2, 322	11.42	.07	$< .001$
	S x G	1, 161	36.23	.18	$< .001$
	S x T	3, 483	33.59	.17	$< .001$
	T x G	3, 483	40.15	.20	$< .001$
	M x T x G	6, 966	16.87	.10	$< .001$
	M x S x G	2, 322	10.09	.06	$< .001$
	M x S x T	6, 966	1.48	.01	$= .18$
	S x T x G	3, 483	3.69	.02	$< .05$

Continued on Next Page...

Table 2.7: Results of the ANOVAs in Experiment 2

	Factor	**df**	**F**	η^2	***p***
	M x S x T x G	6, 966	2.64	.02	$< .05$
d)	Measurement (M)	2, 322	152.62	.49	$< .001$
	View (V)	1, 161	1849.85	.92	$< .001$
	Threat Category (T)	3, 483	216.74	.57	$< .001$
	Group (G)	1, 161	70.32	.30	$< .001$
	M x G	2, 322	80.05	.33	$< .001$
	M x T	6, 966	26.57	.14	$< .001$
	M x V	2, 322	2.99	.02	$= .05$
	V x G	1, 161	0.62	.00	$= .43$
	V x T	3, 483	288.98	.64	$< .001$
	T x G	3, 483	34.91	.18	$< .001$
	M x T x G	6, 966	14.95	.09	$< .001$
	M x V x G	2, 322	1.21	.01	$= .30$
	M x V x T	6, 966	2.82	.02	$< .05$
	V x T x G	3, 483	1.69	.01	$= .17$
	M x V x T x G	6, 966	1.89	.01	$= .08$

As shown in Table 2.8, *t*-tests between the first and second measurement revealed significant training effects for the XRT training group for all threat categories with large effect sizes (all $d > .0.80$). In contrast to Experiment 1, there were also significant effects for the control group, although with rather low effect sizes (all $d < 0.56$). Thus the conventional CBT used in Experiment 2 also resulted in performance increases although much less than XRT.

By an ANOVA with measurement and set as within-participants factors and group as between-participants factor, we investigated if training effects can also be shown for threat objects which were not included in the training sessions. There were main effects and interactions for all factors showing similar results as in Experiment 1 (see Table 2.7b for details). As in Experiment 1, a large transfer effect was found

Table 2.8. Results of the t-tests comparing the categories between first (t1), second (t2), and third (t3) measurement

XRT training group	***t***	**df**	***p***	***d***
Guns t1 - t2	-6.01	83	< .001	0.86
IEDs t1 - t2	-12.84	83	< .001	1.74
Knives t1 - t2	-5.81	83	< .001	0.80
Other t1 - t2	-12.30	83	< .001	1.64
Guns t1 - t3	-8.19	83	< .001	1.15
IEDs t1 - t3	-20.22	83	< .001	2.70
Knives t1 - t3	-10.97	83	< .001	1.48
Other t1 - t3	-16.46	83	< .001	2.18
Control group	***t***	**df**	***p***	***d***
Guns t1 - t2	-2.19	78	< .05	0.23
IEDs t1 - t2	-3.60	78	< .01	0.42
Knives t1 - t2	-2.73	78	< .01	0.33
Other t1 - t2	-1.46	78	<¡ .15	0.18
Guns t1 - t3	-2.72	78	< .01	0.34
IEDs t1 - t3	-4.61	78	< .001	0.56
Knives t1 - t3	-2.05	78	< .05	0.23
Other t1 - t3	-2.59	78	< .05	0.30

(see Figure 2.12). Not only for the prohibited items of set A, which were included in the training library of XRT, but also for the prohibited objects not used in training of set B, screeners of the XRT training group showed a large increase in detection performance after training. Paired-samples t-tests between the first and second measurement showed training effects for both sets and also for both groups whereas again large effect sizes were found for the XRT training group and small effect sizes for the control group (trained group set A: $t(83) = -13.10, p < .001, d = 1.77$ and set B: $t(83) = -9.53, p < .001, d = 1.24$, control group set A: $t(78) = -2.32, p < .05, d = 0.24$ and set B: $t(78) = -3.00, p < .01, d = 0.32$). Pairwise t-tests showed no significant difference in the difficulty of set A and Set B for both groups at the first measurement (XRT training group: $t(83) = 1.16, p = .25, d = 0.10$, control group: $t(78) = 1.93, p = .06, d = 0.19$).

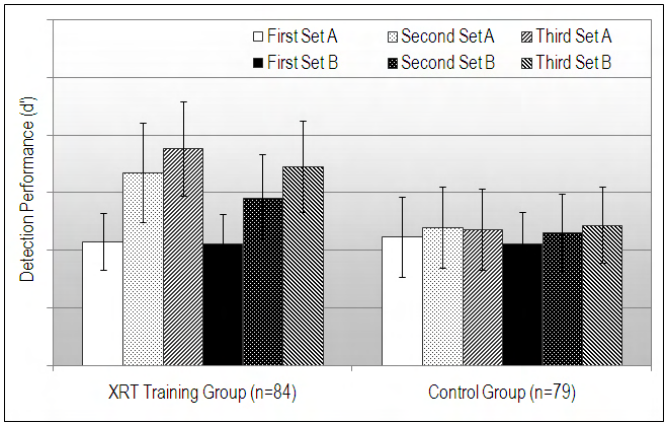


Fig. 2.12. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement for set A and set B separately.

Figure 2.13 includes also the threat category in the analysis. Paired samples t-tests were calculated in order to investigate if the training effect between the first and second measurement was significant for each category in both sets for the XRT training group. Results revealed significant effects for all categories in each set ($p < .01, d = 0.51$ for knives in Set B, $p < .001, d > 0.74$ for all other categories). Thus, as in Experiment 1, XRT resulted in large detection performance increases even for prohibited objects that are not part of the XRT image library (X-Ray CAT image set B). For the control group the difference between the first and third measurement was calculated in order to maximize the chances for finding a significant training effect. The following t-tests were significant: IEDs for both sets, knives only for set A, and other threat objects for both sets ($p < .05, d > 0.23$). All other values were not significant ($p > .06, d < 0.28$) and reveal no effect of training between the different measurements.

As in Experiment 1, individual $d´$ scores were subjected to an extended ANOVA with the within-participants factors measurement, X-Ray CAT image set, threat category, and the between-participants factor group. All main effects and interactions were significant except the interaction between measurement, set, and threat category (see Table 2.7c for details). In contrast to Experiment 1, the ANOVA revealed a main effect of set and significant interactions with set. However, as can be seen in Figure 2.13 they were rather small, which implies large transfer effects. As in Experiment 1, the results clearly show a training effect for each category and in both sets. This is consistent with the results of the t-tests explained above. The training

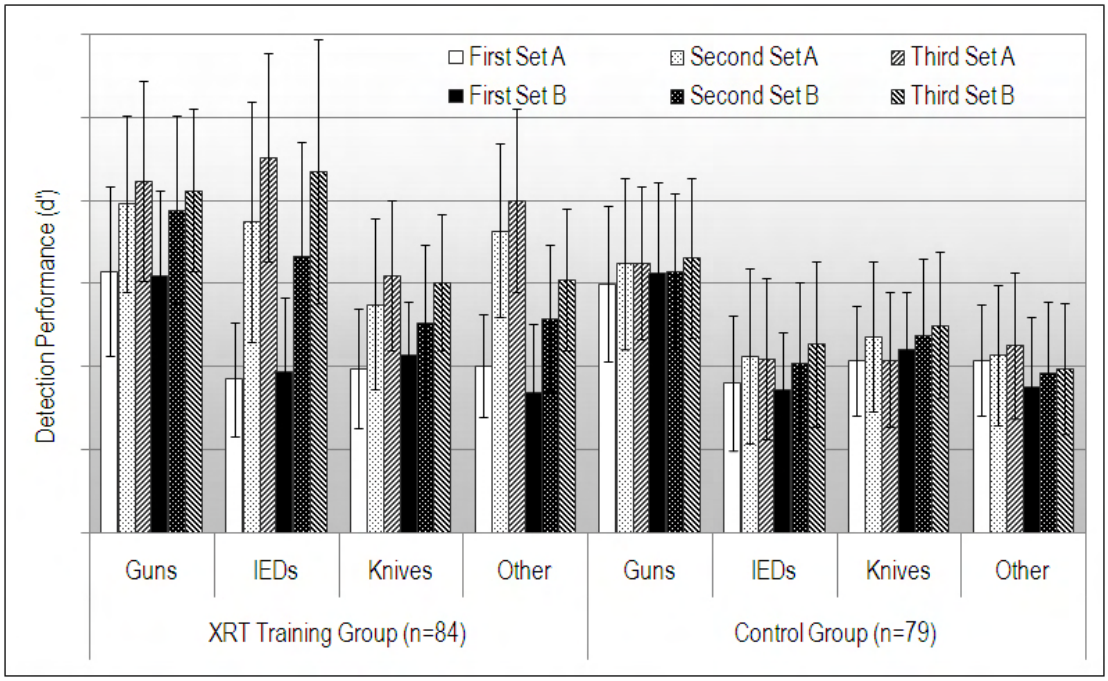


Fig. 2.13. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B and each threat category separately.

effect that was found for the control group revealed itself also in the sets, that is, there was a transfer effect for the control group, too. Last, the effect of viewpoint was investigated calculating a four-way ANOVA. Results show clear main effects of measurement, view, threat category, and group. For details on interactions please refer to Table 2.7d. Detection performance is clearly much higher for objects that are shown in the easy view (View 1) than for the objects that are shown from an unusual viewpoint (see Figure 2.14).

This effect is valid for all threat categories and for the XRT training group as well as for the control group. However, the viewpoint effect is not the same for different threat categories. The graphs in Figure 2.14 suggest that the largest viewpoint effect can be observed for the detection of knives, the smallest one for IEDs. As in Experiment 1, pairwise t-tests showed a significant increase in detection performance at the second measurement for both views for the XRT training group for all four threat categories ($p < .01, d > 0.49$. For the easy view, the control group showed a significant effect for IEDs only ($p < .05, d = 0.32$), all other t-tests were not

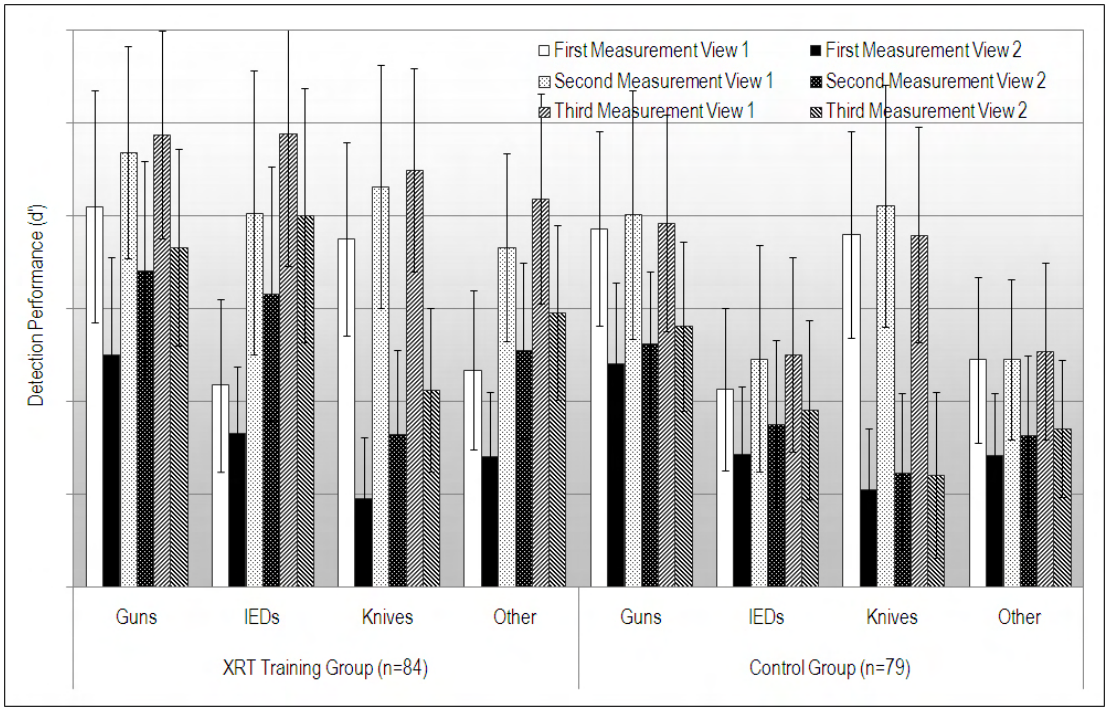


Fig. 2.14. Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second, and third measurement for both views and each threat category separately.

significant ($p > .07, d < 0.25$). For the difficult view all t-test with one exception were significant for the control group ($p < .05, d > 0.26$). Only the training effect of knives in the rotated view was not significant $p = .07, d = 0.24$ (see Table 2.9 for details). But the results show that although some significant effects in the control group were observed, effect sizes were small compared to those of the XRT training group.

In summary, very similar results as in Experiment 1 have been found in Experiment 2. A large and significant training effect was observed for the group who trained with XRT compared to a control group who used a conventional CBT for the same time. A significant training effect has been observed for all four categories (guns, knives, IEDs, and other) for the XRT training group, whereas the effect size varied between categories. Also a large transfer of the acquired knowledge about the visual appearance of objects used in the training (set A) to similar looking objects not used in training (set B) was found for the XRT training group. Additionally, a viewpoint effect could be observed which shows that unusual views of forbidden

Table 2.9. Results of the t-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	$t(83)$	p	d
Guns: V1t1 - V1t2	-3.59	$< .01$	0.49
IEDs: V1t1 - V1t2	-10.93	$< .001$	1.51
Knives: V1t1 - V1t2	-4.35	$< .001$	0.48
Other: V1t1 - V1t2	-9.79	$< .001$	1.42
Guns: V2t1 - V2t2	-5.46	$< .001$	0.82
IEDs: V2t1 - V2t2	-9.99	$< .001$	1.45
Knives: V2t1 - V2t2	-5.79	$< .001$	0.88
Other: V2t1 - V2t2	-10.33	$< .001$	1.40
Control group	$t(78)$	p	d
Guns: V1t1 - V1t2	-1.07	$= .29$	0.13
IEDs: V1t1 - V1t2	-2.64	$< .05$	0.32
Knives: V1t1 - V1t2	-1.87	$= .07$	0.25
Other: V1t1 - V1t2	-0.05	$=.96$	0.01
Guns: V2t1 - V2t2	-2.35	$< .05$	0.26
IEDs: V2t1 - V2t2	-3.24	$< .01$	0.41
Knives: V2t1 - V2t2	-1.81	$= .07$	0.24
Other: V2t1 - V2t2	-2.11	$< .05$	0.28

objects are much harder to detect than canonical views. In contrast to Experiment 1, the control group also showed increases of detection performance, which implies that the conventional CBT used in Experiment 2 is more effective than the one used in Experiment 1 (although still much less effective than XRT). Moreover, there was also a transfer effect for the control group.

2.4 General Discussion

The first aim of this study was to investigate how well airport security screeners can detect guns, knives, IEDs, and other prohibited items in X-ray images of passenger bags. Two experiments conducted at two European airports provided very similar results. A computer-based test (X-Ray CAT) was conducted before and after three and six months of weekly (about 20 min per screener) CBT at each airport. The

first measurement revealed that guns were detected best, followed by knives, other prohibited items, and IEDs. In both experiments and airports, one group used an adaptive CBT (X-Ray Tutor, XRT) with individually adaptive algorithms, a large library of prohibited items depicted in a variety of different views, and automatically created prohibited item to bag combinations (see Schwaninger, 2004, for details). The other group used a conventional CBT system with no adaptive algorithms, a smaller image library, and fixed combinations of threat items in bags. While XRT was used in both experiments and airports, two different conventional CBT systems were used for the control groups of Experiment 1 (airport 1) and Experiment 2 (airport 2). At both airports, XRT training group results revealed a training effect for all types of threat objects (guns, knives, IEDs, and other prohibited items). However, effect sizes differed remarkably for the four categories. While guns were detected best and IEDs were detected worst at the beginning, IED detection of the XRT training group was as good as or even slightly better than gun detection after several months of training. This shows that the detection of IEDs is not difficult per se, but rather depends on the training of screeners. Note that all IEDs used in this study contained a detonator, wires, explosive, a triggering device, and a power source. Therefore, these conclusions are only applicable to the detection of multi-component IEDs. However, a large training effect for IEDs can be expected because they are usually not encountered at airport security checkpoints and therefore not known to screeners without enhanced training in IED detection. The relatively large training effect for the category "other" which includes self defense gas spray, electric shock devices etc. might also be explained by less on the job exposure of these prohibited items. In a study with hold baggage screeners, large training effects for IEDs were also found, which is consistent with results of this study (Schwaninger & Hofer, 2004). In contrast to IEDs and other prohibited items, guns seem to be well known by screeners either because of their typical shape or the frequency by which they are encountered at the airport security screening checkpoint (e.g., toy guns). Therefore, detection performance before training is already high for guns and a large improvement is impossible. It is also noticeable that detection for knives showed the smallest training effect in both experiments. Although the detection

was at the baseline measurement higher than for IEDs and other prohibited items, after six months of training screeners' performance was poorest for knives. On average, knives are smaller than IEDs and other threat items and show less diagnostic features. This might be a reason for the lower detection performance increase for this threat category. While training with XRT resulted in large training effects, the tested conventional CBT systems were less effective. In Experiment 1, there were no training effects at all, while only small training effects were observed for the conventional CBT system used in Experiment 2. This could be due to one or a combination of the following reasons: First, the conventional CBT systems tested in this study do not feature individually adaptive training algorithms like XRT (see Schwaninger, 2004, for details). Second, in contrast to XRT, the conventional CBT systems did not contain such a large image library with many prohibited items depicted from a variety of different viewpoints. Third, while in XRT prohibited items are blended into X-ray images of passenger bags on the fly using scientifically validated and individually adaptive algorithms based on image measurement, as described in Schwaninger, Michel, and Bolfing (2007). The conventional CBT systems used in Experiment 1 and 2 have only fixed combinations of prohibited items in bags. Finally, we had to rely on the statement of the appropriate authority and the security companies regarding the amount of training that was conducted by screeners of the control group and the XRT training group, which should have been on average 20 min per week per screener. Analysis of XRT training data showed, that this was clearly fulfilled for screeners of the XRT training group at both airports. The X-Ray CAT is composed of two comparable (similar looking) sets (set A and set B) and only the threat objects of set A were included into the XRT training system. Therefore, transfer effects can be tested, that is, whether training with certain prohibited items helps increasing detection of other prohibited items that are not contained in the training. Overall, the comparison of the two sets A and B at the baseline measurement (before training) shows no significant difference. However, in Experiment 1 there was a slight difference for the control group between the two sets indicating that the two sets are not exactly equal in terms of image difficulty for this sample. But this possible objection to the transfer effect can be disproved with

two arguments: first, the effect size was only small according to the conventions of Cohen (1988) and second, only one of the two control groups showed a significant difference. Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only. The small training effect for the control group in Experiment 2 is also reflected in the detection increase of both sets after training. Although the conventional CBT system of this control group did not contain any objects from the test, the training with this training system apparently also led to a transfer of knowledge to the detection of objects in the test. In another study it would be interesting to compare the objects that are used in the two training systems used by the control groups regarding their similarity to the test objects. Contrary to our results, Smith, Redford, Gent, and Washburn (2005) found a large decrease in screeners' detection performance when specific objects used in the training were replaced with new images belonging to the same categories (see also Smith, Redford, Washburn, & Taglialatela, 2005). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in X-ray image interpretation. However, it might be possible that the objects of the set not used in training in our study are so similar to the objects used in training that a specific-token familiarity led to the detection performance increase and not a true generalization effect. The observation that a transfer effect in knives was lacking would mean that the objects in set A and set B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer's (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it would mean that those objects were very similar in order to create a specific-token familiarity and therefore a training effect. In both Experiments a large viewpoint effect was also revealed. This is consistent with view-based theories of object recognition (for reviews

see for example Graf et al., 2002; Hayward, 2003; Tarr & Bülthoff, 1995, 1998). After training, easy and difficult views were much better recognized. Interestingly, there was no significant interaction between measurement and viewpoint, that is, although training resulted in improved performance for difficult views, the viewpoint effect (impairment for unusual vs. canonical views) remained stable even after six months of training. However, it must be pointed out that the XRT training algorithm only provides the screeners with unusual views of objects once a screener can detect a prohibited item well when depicted from an easy perspective. That is, when screeners start to train with XRT all threat objects are shown in easy views. Only if these objects are detected reliably, the difficulty level is increased for a certain threat item by showing it in more difficult views (Schwaninger, 2004). Thus, it is unclear whether a significant interaction between viewpoint and measurement would have been observed if the training duration would have been increased (e.g., to one year). In conclusion, it stands to reason that recognition of forbidden objects in X-ray images is dependent on exposure and this has very important implications for an adaptive training system. It has been assumed that different views of each object become associated with one another during object rotation, either through active learning or through passive experiencing of the successive appearance of nearby views (Földiák, 1991; Stryker, 1991). Hence, it is important that during training screeners are getting feedback on which forbidden object has been detected or missed. This feedback shows the photograph and also the X-ray image of that forbidden object always in the canonical view whereas the forbidden object merged into a bag is presented in different viewpoints. This leads to an association between an unusual view of an object and the canonical view which results in a sequential pairing of these views with each other (G. Wang, Obama, Yamashita, Sugihara, & Tanaka, 2005). This association, which forms during learning, is thought to underlie object recognition ability across changes in viewing angle (Palmeri & Gauthier, 2004). For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks & Belcourt, 2006). In any case, our findings show that the

knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects and also the effect that rotated views are much harder to detect can be decreased with training. To make sure that objects are well detected it is important that a large and representative image library of prohibited objects is used and that these objects are learned from different viewpoints. Additionally, the library should be updated constantly to adapt to new threats. Overall, this study has shown that adaptive CBT can be a powerful tool to increase screeners' X-ray image interpretation competency in an efficient and effective way.

Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). Many thanks to Zurich State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

3

Change of Search Time and Non-search Time due to Training in X-ray Baggage Screening

3.1 Introduction

Threat detection using X-ray images in airport security screening is a process that only recently has become a major interest in research concerning object recognition and inspection. The task of airport security screeners is to recognize threat objects of various categories (guns, knives, improvised explosive devices, etc.) in passenger baggage. By applying findings about object recognition this important part of common airport security concepts can be improved. The knowledge about how objects are perceived has allowed the creation of a computer-based training system X-Ray Tutor (XRT) (Schwaninger, 2005b). This training system considers the factors influencing the recognition of objects, which are: the viewpoint in which an object is depicted; the superposition of the object by other objects in the bag; and the number and type of other objects in the bag (Schwaninger, 2003b; Schwaninger et al., 2004; Wallis & Bülthoff, 1999). XRT is individually adaptive. It starts with threat items depicted in easy views and increases image difficulty for each individual trainee by showing threat items in more difficult views and in more complex bags and with increasing superposition by other objects. In order to prevent screeners from memorizing images of bags, combinations of images of bags and threat objects are created at the point of use. This approach considers the individual training level and visual-cognitive abilities of each screener. Security inspection is a form of visual inspection but there exist few studies quantifying human performance of security screening (see Gale, Mugglestone, Purdy, & McClumpha, 2000; Gale, Purdy, & Wooding, 2005; Liu,

Gale, Purdy, & Song, 2006; McCarley et al., 2004; Schwaninger et al., 2004). The last two have shown that training increases the threat detection performance of airport security screeners significantly. A deeper comprehension of the effect of training could be gained if the specific task of security inspection could be compared to more general models of industrial and other inspection tasks. Recent findings confirmed the applicability of a two-component model of visual inspection (Drury, 1975; Spitz & Drury, 1978) to X-ray screening data (Ghylin, Drury, & Schwaninger, 2006). Spitz and Drury (1978) assumed the inspection task was composed of search and decision components. Each of these components occupies part of the time needed for completing the task. Using the equations formulated by Drury (1975) and Spitz and Drury (1978) the total inspection time can be divided into the functional components of search and decision time. The general model created and tested by Drury (1975) and Spitz and Drury (1978) is:

$$\text{P(true target)} = \left[1 - \exp\left(-\frac{(t - NST_{hit})}{ST_{hit}}\right)\right] * Pd_{hit} \tag{3.1}$$

$$\text{P(false alarm)} = \left[1 - \exp\left(-\frac{(t - NST_{FA})}{ST_{FA}}\right)\right] * Pd_{FA} \tag{3.2}$$

Where:

P(true target) = detect a true target at or before time t

P(false alarm) = make a false alarm at or before time t

ST_{hit} or ST_{FA} = search time for hits or false alarms

NST_{hit} or NST_{FA} = non-search time for hits or false alarms

Pd_{hit} or Pd_{FA} = the probability of detection found from the raw data for hits or false alarms

t = the various raw reaction times obtained from the data

The model assumes an approximately exponential relationship between the time needed for searching a target and the cumulative probability of detecting a target (Morawski, Drury, & Karwan, 1980). Search time includes the visual scanning of an area to be searched (i.e., eye movements) and is terminated by either directing the

attention to a suspicious part of this area (i.e., potential threat object in this case) or by deciding to stop searching. Decision time is everything except search and is more correctly called non-search time. It includes, among other things, the fixation of the suspicious object, the matching of the visual stimulus with representations stored in the visual memory, the decision (i.e., actually is threat object or not), and the time to execute the response. This model has been applied successfully to a number of different screening data sets (e.g., Ghylin et al., 2006). Applying this two-component model of visual inspection helps in identifying the sub processes of the whole inspection task and therefore may give some evidence about how the two processes improve differentially due to training. Feature Integration Theory (FIT) (Treisman & Gelade, 1980) assumes that visual features of objects are represented in feature maps. Features are those stimulus attributes that are processed rapidly and in parallel across the field of view. As soon as a visual field of an observer contains more than one object the binding problem arises (Treisman, 1998). Features of various objects have to be combined correctly and assigned to the right object in order to perceive it correctly. In the original feature integration model (Treisman & Gelade, 1980) search for feature conjunctions is not allowed. Wolfe (1994) found that combination of feature information permits the efficient, guided search for feature conjunctions and postulated this in the *Guided Search 2.0* model of visual search. When a threat object is, presumably deliberately, stowed in a bag, it is typically not just a target among several distractors. Most likely its shape on the X-ray image is interrupted by other objects surrounding it and superimposed on it. This complicates the assignment of features to an object, particularly if this object is only poorly known. If airport security screening training for threat object detection has the effect of creating internal representations of objects used in the training and storing them and making them available, respectively, in the visual memory, then detection should improve because features are known and recognized better. We would also expect that, with growing knowledge about the visual appearance of threat objects in X-ray images of passenger bags, the number of required features for the recognition of an object can be limited or once separately perceived features can be combined as belonging to one object and thus becoming one feature.

This would require building new feature maps. Considering the assumption of FIT that visual search for a combination of features is serial and therefore more time consuming than the visual search for a unique feature, the assumption would be that detection time would decrease for threat objects that are detected better. In other words, with increasing detection performance due to training the detection time, more explicitly the search time, should decrease. Ghylin et al. (2006) found an enhancement of both the search process and the non-search process of inspection in the search for improvised explosive devices (IEDs). This study extends the analysis to three other threat categories (guns, knives, and other threat objects), potentially validating the findings of Ghylin et al. (2006) for other objects than IEDs. Prior inspection research has found individual differences among inspectors to be large (Czaja & Drury, 1981a; Dollinger & Hoyer, 1996; McPhee, Scialfa, Dennis, Ho, & Caird, 2004; Riegelnig & Schwaninger, 2006; Schwaninger et al., 2004; M. J. J. Wang, Lin, & Drury, 1997) so that analyses of detection performance and reaction times in this study are controlled for age, gender and on the job experience. There has been little examination of screener demographics in relation to either overall performance parameters or search and non-search measures. In the broader inspection literature, age, gender, and experience have received some attention. Older screeners tend to work more slowly and at times have lower detection performance (Czaja & Drury, 1981a; McPhee et al., 2004), although any deficits can be largely negated by age-specific training (Czaja & Drury, 1981b). The mechanisms for age-related deficits are quite well understood (e.g., Fozard, 1990). Aging decreases pupil diameter, spatial resolution, visual acuity (particularly dynamic), contrast sensitivity (Owsley, Sekuler, & Siemsen, 1983), depth perception and visual search (e.g., Plude & Hoyer, 1986) but not colour vision or temporal resolution. Gender has not been found to be related to inspection performance (e.g., M. J. J. Wang & Drury, 1989). Experience can either refer to novice/expert differences or to the effective length of experience of those with expertise. For example, Dollinger and Hoyer (1996) found novice/expert differences while Leach and Morris (1998) found no effect of longer experience. These findings are typical of experience results: Novices differ considerably from experts, but length of time on the job beyond initial training may not show much effect.

3.2 Methods

3.2.1 Participants

A total of 193 airport security screeners of a European airport, all with on-the-job experience of airport X-ray screening were used. Of these 193 screeners, 98 (44 females, mean age 36.3 years, mean time on job 3.0 years; 54 males, mean age 40.0 years, mean time on job 3.0 years) of them trained for six months with X-Ray Tutor while the other 95 (48 females, mean age 35.1 years, mean time on job 3.0 years; 47 males, mean age 36.9 years, mean time on job 3.3 years) received no training with X-Ray Tutor during this period. All then took the X-Ray Competency Assessment Test (X-Ray CAT).

3.2.2 Materials and Procedure

The X-Ray CAT is composed of 128 X-ray images of bags. Each image can include one threat object out of four threat categories according to international threat image projection system specification (guns, improvised explosive devices IEDs, knives, and other threat objects). Stimuli were created from Smiths-Heimann Hi-Scan 6040i colour X-ray images. Each bag was used once containing a threat object and once containing no threat object, giving a total of 256 images. In each threat category 16 objects were depicted once in an easy viewpoint (frontal view) and once in a more difficult rotated viewpoint (85 degrees vertical or horizontal rotation). All threat objects were combined with bag images so as to ensure that for each threat object the degree of superposition (i.e., how much the threat object is superimposed by other objects in the bag) was equal for both viewpoints. The difficulty of the bags was equated across all categories and viewpoints by matching prior data on False Alarm Rates for each bag image. For a more detailed description of X-Ray CAT see Koller and Schwaninger (2006). The X-Ray CAT is a component of the X-Ray Tutor training system and can be programmed to appear anytime when a screener is training. The only visible difference for the screener between test and training was that feedback did not appear during the CAT test. The appearance of the images was

the same for both training and test and therefore no instruction was needed. The images of the test disappeared after 15 seconds. Screeners had to decide whether the X-ray images contained a threat object or not (NOT OK or OK response). Difficulty ratings had to be provided by changing the position of a slider on a 100 point scale. Response times for each image were measured.

3.2.3 Results

The X-Ray Tutor program measured the response as hit, miss, false alarm (FA), or correct rejection (CR) with the reaction time for each test image and each participant. Detection performance in terms of A' (Pollack & Norman, 1964) was calculated for each threat category separately (see Figure 3.1), using the following formula (Grier, 1971): $A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)]$, where H is the hit rate and F the false alarm rate. If the false alarm rate is greater than the hit rate the equation must be modified (Aaronson & Watts, 1987): $A' = 0.5 - [(F - H)(1 + F - H)]/[4F(1 - H)]$. A' scores were subjected to a univariate analysis of covariance (ANCOVA) with age and years on job as covariates, threat type treated as within-participants factor (guns, IEDs, knives, other) and gender and training (trained vs. untrained group) as between-participants factors. The results are summarized in Table 3.1.

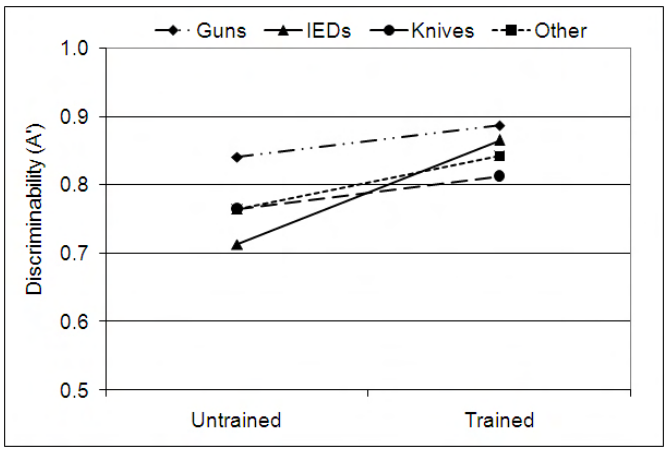


Fig. 3.1. Detection Performance A' for trained and untrained participants for guns, IEDs, knives, and other threat objects. (Note: Performance values are multiplied by an arbitrary constant for security purposes).

A' scores were also subjected to separate univariate ANCOVAs for each threat (see Table 3.2).

Table 3.1. Results of univariate ANCOVAs on A', hit rate (HR), and false alarm rate (FAR)

	A'				HR				FAR			
Factor	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
Threat Type (TT)	-	-	-	-	3, 561	7.12	0.04	< 0.001	n.a.	n.a.	n.a.	n.a.
Training (T)	1, 187	99.03	0.35	< 0.001	1, 187	31.51	0.14	< 0.001	1, 187	12.03	0.06	< 0.001
Gender (G)	1, 187	11.44	0.06	< 0.001	-	-	-	-	1, 187	6.25	0.03	< 0.05
TxTT	3, 561	52.55	0.22	< 0.001	3, 561	70.87	0.28	< 0.001	n.a.	n.a.	n.a.	n.a.
TxG	1, 187	3.99	0.02	< 0.05	-	-	-	-	-	-	-	-
TTxG	-	-	-	-	-	-	-	-	n.a.	n.a.	n.a.	n.a.
Age (A)	1, 187	40.25	0.18	< 0.001	1, 187	10.97	0.06	< 0.01	1, 187	4.97	0.03	< 0.05
Years on Job (Y)	1, 187	7.91	0.04	< 0.01	1, 187	7.31	0.04	< 0.01	1, 187	15.74	0.08	< 0.001
TxA	-	-	-	-	-	-	-	-	-	-	-	-
TxY	-	-	-	-	-	-	-	-	-	-	-	-
TTxA	3, 561	7.21	0.04	< 0.001	3, 561	8.47	0.04	< 0.001	n.a.	n.a.	n.a.	n.a.
TTxY	3, 561	5.16	0.03	< 0.01	3, 561	5.77	0.03	< 0.01	n.a.	n.a.	n.a.	n.a.

Table 3.2. Results of univariate ANCOVAs on A', hit rate, and search and non-search time for the threat categories separately

		Guns				Knives				IEDs				Others			
A'	Factor	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
	Training (T)	1, 187	39.68	0.18	< 0.001	1, 187	24.42	0.12	< 0.001	1, 187	161.49	0.46	< 0.001	1, 187	51.38	0.22	< 0.001
	Gender (G)	1, 187	5.39	0.03	< 0.05	1, 187	5.74	0.03	< 0.05	1, 187	11.47	0.06	< 0.001	1, 187	7.37	0.04	< 0.01
	Age (A)	1, 187	20.01	0.10	< 0.001	1, 187	17.86	0.09	< 0.001	1, 187	37.05	0.17	< 0.001	1, 187	30.67	0.14	< 0.001
	Years on Job (Y)	1, 187	13.01	0.07	< 0.001	1, 187	12.79	0.06	< 0.001	-	-	-	-	1, 187	7.93	0.04	< 0.01
Hit Rate	Factor	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
	Training (T)	1, 187	6.37	0.03	< 0.05	-	-	-	-	1, 187	77.85	0.29	< 0.001	1, 187	29.45	0.14	< 0.001
	Gender (G)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Age (A)	1, 187	3.97	0.02	< 0.05	-	-	-	-	1, 187	14.35	0.07	< 0.001	1, 187	12.80	0.06	< 0.001
	Years on Job (Y)	-	-	-	-	-	-	-	-	1, 187	10.37	0.05	< 0.01	1, 187	5.70	0.03	< 0.05
Search Time	Factor	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
	Training (T)	-	-	-	-	-	-	-	-	1, 186	20.91	0.1	< 0.001	1, 187	4.33	0.02	< 0.05
	Gender (G)	1, 187	14.44	0.07	< 0.001	-	-	-	-	1, 186	5.43	0.03	< 0.05	1, 187	14.19	0.07	< 0.001
	Age (A)	1, 187	16.8	0.08	< 0.001	1, 187	12.55	0.06	< 0.001	1, 186	15.62	0.08	< 0.001	1, 187	19.41	0.09	< 0.001
	Years on Job (Y)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Non-Search Time	Factor	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
	Training (T)	1, 187	48.6	0.21	< 0.001	1, 187	45.47	0.2	< 0.001	1, 186	66.75	0.26	< 0.001	1, 187	78.69	0.3	< 0.001
	Gender (G)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Age (A)	1, 187	9.04	0.05	< 0.01	1, 187	31.99	0.15	< 0.001	1, 186	32.15	0.15	< 0.001	1, 187	17.92	0.09	< 0.001
	Years on Job (Y)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Additionally, hit rates were subjected to a univariate ANCOVA with gender and training as between-participants factors, threat as within-participants factor, and age and years on job as covariates. See Table 3.1 for details. Separate ANCOVAs on hit rates for each threat category were performed (see Table 3.2). An ANCOVA on false alarm values with gender and training as between-participants factors and age and years on job as covariates shows a significant main effect of training (see Table 3.1 for details and values on covariate effects).

The effect of training in detection performance comprised an increase in the hit rate

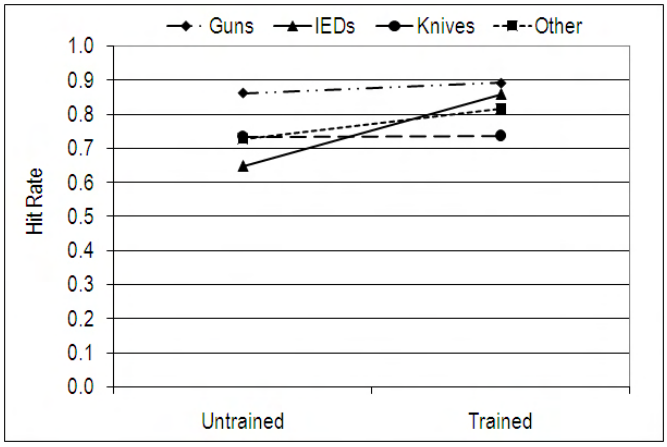


Fig. 3.2. Hit rate for trained and untrained participants for guns, IEDs, knives, and other threat objects (Note: Performance values are multiplied by an arbitrary constant for security purposes).

and a decrease in the false alarm rate. Figure 3.2 displays the hit rate for all threat categories for trained and untrained participants. The pattern is very similar to the one for the detection performance A' (see Figure 3.1). This could be an indication that the difference of the detection performance between trained and untrained participants can mainly be attributed to a change in the hit rate. Nevertheless, the ANCOVA on false alarm values showed a significant effect of training, which means that also the false alarm rate is affected by training. The average reaction time in seconds was calculated for hits, false alarms, misses, and correct rejections for trained and untrained screeners (see Figure 3.3). Reaction times were subjected to an ANCOVA with age and years on job as covariates, gender and training as between-participants factors and answer as a within-participants factor (hit, false alarm, miss, correct rejection). The results including the significance values can be seen in Table 3.3.

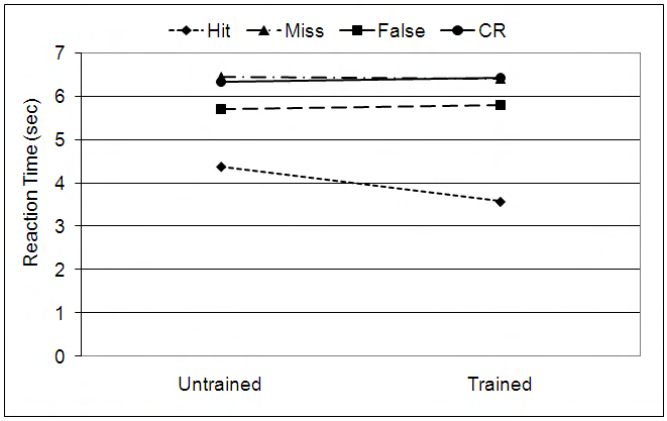


Fig. 3.3. Reaction times in seconds for trained and untrained participants for hits, misses, false alarms, and correct rejections.

	RT			
Factor	df	F	η^2	p
Answer(S)	3, 561	15.51	0.08	< 0.001
Training (T)	-	-	-	-
TxS	3, 561	12.36	0.06	< 0.001
Age (A)	1, 187	18.87	0.09	< 0.001
Gender (G)	1, 187	11.9	0.06	< 0.001
Years on Job (Y)	-	-	-	-
TxA	-	-	-	-
TxG	-	-	-	-
TxY	-	-	-	-
SxA	-	-	-	-
SxG	3, 561	3.72	0.02	< 0.05
SxY	-	-	-	-

Table 3.3. Results of univariate ANCOVAs on reaction time (RT)

Reaction times were subjected to separate univariate ANCOVAs for hits, false alarms, misses, and correct rejections with gender and training as between-participants factors (detailed results are summarized in Table 3.4).

Table 3.4. Results of univariate ANCOVAs on reaction time (RT) for each answer category separately

	Hits				Misses				False Alarms				Correct Rejections			
	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p	df	F	η^2	p
Training (T)	1, 187	38.05	0.17	< 0.001	-	-	-	-	-	-	-	-	-	-	-	-
Age (A)	1, 187	33.90	0.15	< 0.001	1, 187	14.08	0.07	< 0.001	1, 187	11.97	0.06	< 0.001	1, 187	14.53	0.07	< 0.001
Gender (G)	1, 187	7.57	0.04	< 0.01	1, 187	10.21	0.05	< 0.01	1, 187	13.18	0.07	< 0.001	1, 187	9.45	0.05	< 0.01
Years on Job (Y)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

The intercorrelation matrix in Table 3.5 shows the correlations between the predictor variables (age, gender, and years on job) and the performance variables (hit rate, false alarm rate, reaction times of hits, false alarms, correct rejections, and misses) as well as intercorrelations between the set of performance measures.

The mean search time and the mean non-search time were calculated for each screener individually and for hits and false alarms separately by applying the inspection model (Spitz & Drury, 1978, see Introduction) to the reaction times. If there were less than five responses (i.e., hit or FA) available for a person or a reaction time exceeded 14 seconds this data was discarded from analysis. Data elimination was effected for 3.7% of the trials (1839 of 49407 cases with RT bigger than 14 seconds). Final sample sizes were n=98 (hit RT) and 97 (FA RT) for trained participants, and n=95 (hit as well as FA RT) for untrained participants. The hits were again

Table 3.5. Intercorrelation matrix with predictor variables (gender, age, years on job) and performance variables (pHit, pFA, RTHit, RTFA, RTCR, RTMiss)

Correlations

		pHit	Gender	pFA	RTHit	RTFA	RTCR	RTMiss	Age	YOJ
pHit	Pearson Correlation	1	0.016	.509(**)	-.216(**)	0.028	.282(**)	.218(**)	-.184(*)	.168(*)
	Sig. (2-tailed)		0.827	0.000	0.003	0.699	0.000	0.002	0.011	0.020
pFA	Pearson Correlation	.509(**)	.153(*)	1	-0.007	-.229(**)	0.130	0.084	0.086	.266(**)
	Sig. (2-tailed)	0.000	0.033		0.926	0.001	0.071	0.246	0.232	0.000
RTHit	Pearson Correlation	-.216(**)	-.186(**)	-0.007	1	.772(**)	.695(**)	.680(**)	.323(**)	-0.004
	Sig. (2-tailed)	0.003	0.009	0.926		0.000	0.000	0.000	0.000	0.953
RTFA	Pearson Correlation	0.028	-.281(**)	-.229(**)	.772(**)	1	.828(**)	.816(**)	.263(**)	-0.068
	Sig. (2-tailed)	0.699	0.000	0.001	0.000		0.000	0.000	0.000	0.347
RTCR	Pearson Correlation	.282(**)	-.257(**)	0.130	.695(**)	.828(**)	1	.936(**)	.270(**)	0.041
	Sig. (2-tailed)	0.000	0.000	0.071	0.000	0.000		0.000	0.000	0.572
RTMiss	Pearson Correlation	.218(**)	-.259(**)	0.084	.680(**)	.816(**)	.936(**)	1	.252(**)	0.066
	Sig. (2-tailed)	0.002	0.000	0.246	0.000	0.000	0.000		0.000	0.360
Age	Pearson Correlation	-.184(*)	-0.140	0.086	.323(**)	.263(**)	.270(**)	.252(**)	1	-0.045
	Sig. (2-tailed)	0.011	0.052	0.232	0.000	0.000	0.000	0.000		0.531
YOJ	Pearson Correlation	.168(*)	-0.030	.266(**)	-0.004	-0.068	0.041	0.066	-0.045	1
	Sig. (2-tailed)	0.020	0.679	0.000	0.953	0.347	0.572	0.360	0.531	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

separated by threat category (guns, IEDs, knives, and other threat items) where for the IEDs one untrained participant achieved less than five hits within 14 seconds. Sample size therefore was 94 for IEDs. For false alarms no threat categories exist. Using the cumulative distributions of reaction times, search time and non-search time were calculated using a linear regression model (Drury, 1975; Spitz & Drury, 1978). Figures 3.4 and 3.5 show search and non-search times for false alarms and for hits per threat category, respectively.

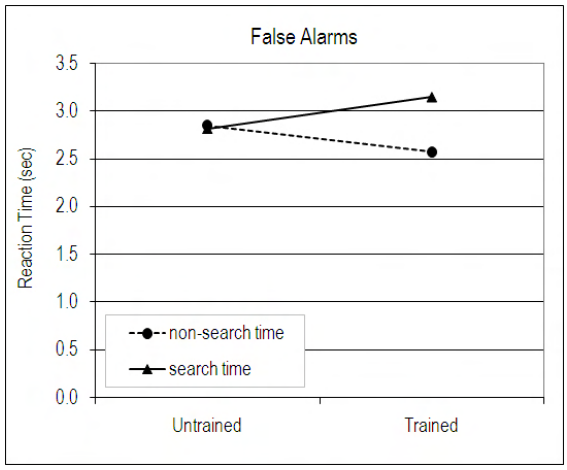


Fig. 3.4. Search time and non-search time for false alarms by trained (n=97) and untrained (n=95) participants.

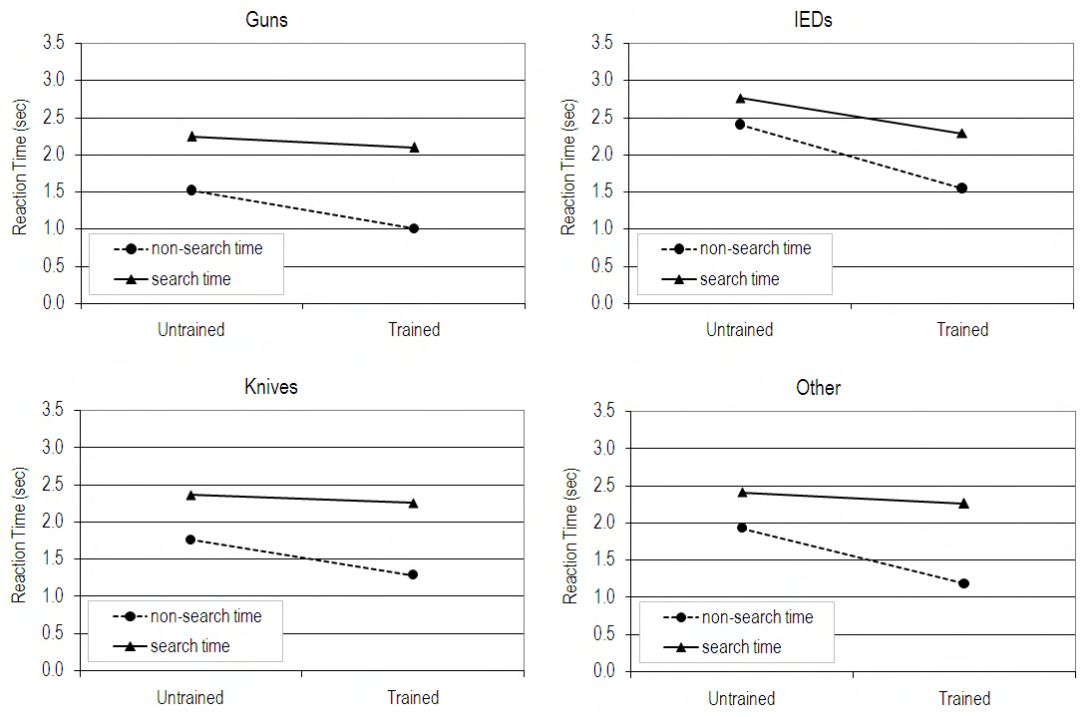


Fig. 3.5. Search time and non-search time for hits, calculated separately for guns, IEDs, knives, and other threat items by trained (n=98) and untrained (n=95; IEDs n=94) participants.

The applicability of the two-component model (Spitz & Drury, 1978) to security inspection task was tested by goodness of fit values (r^2) averaging above 0.9 both for hits ($Mean = 0.961$, $SD = 0.035$) and false alarms ($Mean = 0.905$, $SD = 0.069$). The scores of search times and non-search times for hits were subjected to separate ANCOVAs with age, gender, and years on job as covariates, threat as a within-participants factor and training as a between-participants factor. Table 3.6 summarizes the results.

Individual search and non-search parameters were subjected to separate univariate ANCOVAs for each threat category with training and gender as independent factors and age and years on job as covariates. Non-search time was affected by training in all threat categories (see Table 3.2 for details). The covariate effect of age was significant for all threat categories, whereas gender and years on job never had covariate effects (see Table 3.2). Table 3.2 also displays the effects on search time. Search time was only affected by gender for the categories guns, IEDs, and other

Table 3.6. Results of univariate ANCOVAs on search times and non-search times

	Search Times				Non-search Times			
Factor	df	F	η^2	p	df	F	η^2	p
Threat Type (TT)	-	-	-	-	-	-	-	-
Training (T)	1, 186	8.59	0.04	< 0.01	1, 186	83.38	0.31	< 0.001
Gender (G)	1, 186	11.10	0.06	< 0.01	-	-	-	-
TxTT	3, 558	5.93	0.03	< 0.001	3, 558	11.95	0.06	< 0.001
TxG	-	-	-	-	-	-	-	-
TTxG	3, 558	3.07	0.02	< 0.05	-	-	-	-
Age (A)	1, 186	21.81	0.11	< 0.001	1, 186	31.14	0.14	< 0.001
Years on Job (Y)	-	-	-	-	-	-	-	-
TxA	-	-	-	-	-	-	-	-
TxY	-	-	-	-	-	-	-	-
TTxA	-	-	-	-	3, 558	9.03	0.05	< 0.001
TTxY	-	-	-	-	-	-	-	-

threat objects with small to medium effect sizes, but not for knives. Age had an effect on all threat categories and years on job on none. Training influenced only the search time for the categories IEDs and other threat objects, but not for guns and knives. Gender had significant effects on A' (also on each threat category separately) and false alarm rate with small to medium effect sizes according to Cohen (1988), but not on the hit rate (all analyses of effect sizes are interpreted according to Cohen, 1988). The effect of training is slightly influenced by gender that means there was a significant interaction between gender and training with a small effect size (see also Figure 3.6). There was a significant effect of gender on the reaction time with a medium effect size (see Tables 3.1 - 3.3).

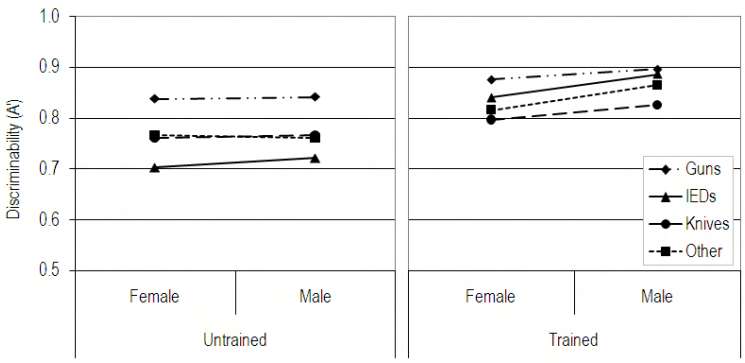


Fig. 3.6. Detection Performance A' for trained and untrained participants for guns, IEDs, knives, and other threat objects for males and females separately. (Note: Performance values are multiplied by an arbitrary constant for security purposes).

Age had significant effects on A' (also on each threat category separately), on the false alarm rate and on the hit rate (all threat categories except knives). The effect size was large for A' and small to medium for hit and false alarm rates. Age had also a significant effect on the reaction time with a medium effect size (see Tables 3.1 - 3.3). Job experience (years on job) had significant effects on A' (all threat categories except IEDs) as well as on hit rate (only on IEDs and other threat objects) and false alarm rate. Effect sizes were small to medium. Reaction time was not affected by working experience (see Tables 3.1 - 3.3).

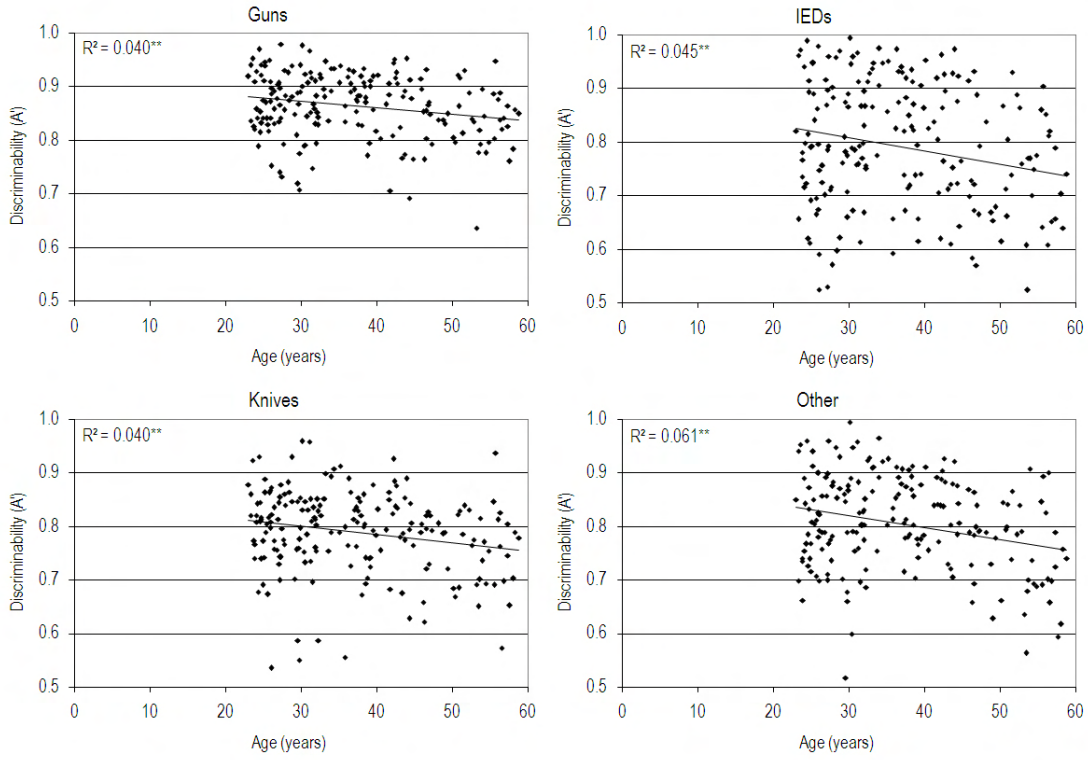


Fig. 3.7. Correlation between age and A' of each threat category. ** = Correlation is significant at the 0.01 level.

Significant correlations were found between the detection performance A' and age and between the detection performance A' and years on job, separated for threat category (see Figures 3.7 and 3.8). However, if controlled for age, the correlation between detection performance A' and job experience (years on job) is not significant anymore ($r = .069, p = .34$).

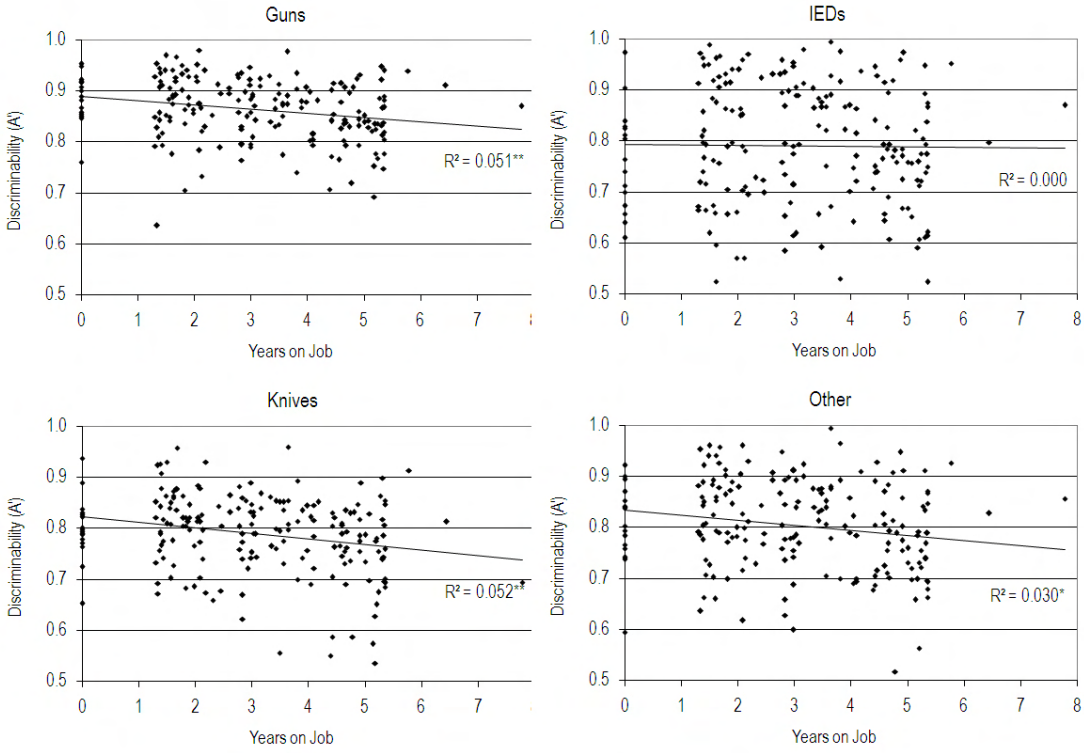


Fig. 3.8. Correlation between years on job and A' of each threat category. ** = Correlation is significant at the 0.01 level, * = Correlation is significant at the 0.05 level.

3.3 Discussion

The data from this study show a substantial increase of threat detection performance in X-ray security screening due to training, regardless of age, gender, or on the job experience, confirming prior findings of Schwaninger and Hofer (2004), McCarley et al. (2004), and also of Ghylin et al. (2006), in a more limited similar study of IEDs. Age and gender were taken into account because there is previous evidence that many aspects of cognition are impaired because of aging (for a review see Craik & Salthouse, 2000) and that gender influences cognitive tasks (see Halpern, 1992, for a review). Although our results point to such differences in detection performance, with partly significant values, the effect sizes are small to medium according to the conventions of Cohen (1988). From our results, detection performance decreases with increasing age and also with increasing job experience. However, since the correlation disappears with a partial correlation, controlling for age, this last interrelation

presumably exists because age and job experience are confounded variables. As for gender, males perform overall slightly better than females. As mentioned before, detection performance increase due to training was not affected by these factors, except for gender, where males tend to benefit a little more than females. Riegelnig and Schwaninger (2006) contributed a more detailed study on the influence of age and gender on detection performance in X-ray screening.

The aim of this study was to analyze the effect of training on a more explicit level, that is, which functions (search, decision) of the threat detection process change because of training. High goodness of fit values ($r^2 > 0.90$) confirm the applicability of the two-component inspection model to X-ray security screening data, as was also found by Ghylin et al. (2006). The application of the two-component inspection model (Spitz & Drury, 1978) to X-ray screening data allows a more detailed investigation of the effect of training on inspection performance of X-ray security screeners. Findings of industrial inspection studies on the cumulative distributions of reaction time for visual search showed that the inspection process can be divided conceptually and operationally into two sub processes. The search process comprises the actual searching of an area (i.e., by a sequence of eye movements); the non-search process comprises all other components of the search (e.g., identification, recognition, decision, response execution, etc.). A similar model has been proposed by Gale et al. (2005) for use on inspection studies. This comprises an initial glance, then serial search followed by "detection and interpretation". Eye movements, rather than cumulative distribution fitting, were proposed to validate that model. Using the two-component inspection model, portions of the reaction time, that is, the time needed for the whole search, can be assigned to one of the two sub processes. Analyzing search and non-search time can give a better understanding of which processes of search change due to training. These findings also quantify the effectiveness of the training system. Comparing the reaction times of trained and untrained screeners reveals a significant decrease for hits but not for false alarms, correct rejections or misses. This, and also the effect of training on hit rates, proves the effectiveness of the training system. Screeners have to learn to detect threat items, and therefore hit rates increase and the time needed for detection decreases, respectively. Although

reaction time for false alarms was not affected significantly by training, false alarm rate actually was. However, as discussed in the next paragraph, reaction time of search and non-search components of false alarms both showed an effect of training. The comparison of the search and the non-search time, respectively, of trained and untrained screeners revealed a significant decrease of the non-search time for hits due to training, specifically to each threat category. Search time of hits was also affected by training. Examining the separate threat categories, this effect was only significant for IEDs and other threat objects, but not for guns and knives. Presumably, guns and knives are well-known objects for experienced screeners so that training does not affect their detection substantially. For false alarms, the tendency for non-search time goes in the same direction but for search time the untrained screeners performed faster. These contrary effects lead to the fact, that overall reaction time for false alarms shows no difference between trained and untrained screeners, as mentioned earlier. The substantial increase of detection performance A' for trained screeners indicates the more effective search behaviour they achieve due to training. As trained screeners also showed significantly lower non-search times than untrained screeners, the training effect seems to come from faster detection, matching of ominous objects seen in the bag with memory representations of prohibited threat objects, and faster recognition and decision about whether the object actually is a threat. Note also that the faster search times of trained screeners for hits were not accompanied by reduced times for correct rejection, that is, the times when the screener failed to find a threat and moved to the next bag image. Thus search was in fact more thorough after training, in that stopping time was a greater multiple of mean search time for hits. An improvement in speed of search is possible if threat objects are learned and therefore stored in the visual memory which indicates that the training system effectively provides more exemplars of threat images. It becomes easier to recognize common attributes of threat objects because they are represented in the visual memory. A decrease of search time implies faster search, but only for IEDs, confirming the findings of Ghylin et al. (2006), and for other threat items. It remains to be investigated why search only gets faster for IEDs, and slightly faster for other threat objects, but not for guns and knives. The former category has wide visual

variability as IEDs can take many forms and their individual components can be scattered throughout the bag. In all studies of security X-ray inspection, they are harder to detect without training (Drury, Ghylin, & Holness, 2006). In contrast the 'other' threat items list changes frequently so that it can never achieve the same level of familiarity as guns or knives, or even IEDs. Thus IEDs have more 'head room' for improved detection and other threat items are always in a learning cycle. It has also to be observed if this speed up continues with training or if it is a side effect of the dramatic improvement of detection performance of IEDs which is far greater than for the other categories. This would imply that search stops improving as soon as detection performance of IEDs is at a higher level for all screeners. A possible reason for the decrease of search time for IEDs and other threat objects could be the building of new feature maps (Treisman & Gelade, 1980) for IEDs and some threat objects belonging to the category of other (e.g., gas spray, tazer, etc.). The assumption is that prior to training they exist with smaller probability than feature maps for guns or knives, because IEDs and other threat objects are rarely to never seen in everyday life, unlike guns or knives, and therefore are mostly unknown to untrained screeners. The applicability of the two-component inspection model (Drury, 1975; Spitz & Drury, 1978) to X-ray inspection provides the possibility of investigating the inspection process and its change due to training more closely and to gain more knowledge about the individual components within inspection. This helps improving the technologies, procedures, and methods currently in use for X-ray security screening, and therefore optimizes the human-system interface.

Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403).

Part II

Benefit or Drawback? Potential Decision Aids for X-Ray Screening

4

Do "Image Enhancement" Functions Really Enhance X-Ray Image Interpretation?

4.1 Introduction

In recent years, the importance of baggage X-ray screening at airports has increased dramatically. The image quality of older X-ray screening equipment was sometimes in need of improvement. For example, an early version of a coloring algorithm as enhancement function did not serve the purpose of increasing detection performance of threat objects, actually it impaired it. This was due to the occlusion of object parts by the opaque coloring algorithm (Schwaninger, 2005a, 2005b). But there has been much technological progress in the last years, especially regarding X-ray screening machines, which nowadays provide high image quality and various image enhancement functions (IEFs). The main objective of such functions is to process an image so that the result is more suitable than the original image for a specific application as for example X-ray screening at airports (Gonzalez & Woods, 2002). In X-ray images, the image enhancements might increase the visibility of objects within the bag and remove background noise. State-of-the-art X-ray machines provide many IEFs. The aim of this study is to investigate whether IEFs actually help human operators (screeners) to better detect threat items in X-ray images of passenger bags. Interestingly, reports regarding an evaluation of IEFs have not been publicly available except two recent publications (Klock, 2005; Schwaninger, 2005b). Klock (2005) examined whether IEFs increase screeners' threat detection performance when visually inspecting carry-on bags using a Rapiscan emulator. She found that high penetration, organic stripping and inorganic stripping functions resulted in decreased probability of detection (see below for more information on different IEFs). Crystal clear, black and white, and low penetration resulted in the best performance, while it should be noted that the original color image was not included in the analysis. Klock (2005) also found that these effects are dependent on threat type, that is, whether guns,

knives, improvised explosive devices (IEDs), or other prohibited items had to be detected. Schwaninger (2005b) reported a study on the effects of IEFs for the detection of IEDs in hold baggage. He found that the original image resulted in the best performance, whereas the organic stripping, organic only and luminance negative functions substantially impaired detection of IEDs. The purpose of this study is to extend previous research in order to evaluate the value of different IEFs. In addition, a comparison between IEFs used in cabin baggage screening (CBS) and hold baggage screening (HBS) was of interest. The nine IEFs examined in this study can be applied to the X-ray images online when working at an aviation security checkpoint. Each pixel in the image format used in these X-ray machines has a material and a luminance value. To show the images on a screen, the pixel values are color coded using red for organic, blue for metallic and green for mixed organic/metallic material. The luminance value defines the luminance of the pixel. Table 4.1 gives an overview and description of all IEFs used in Experiment 1.

Table 4.1: Image Enhancement Filters

Grayscale (GR)	The Grayscale filter removes the material information from the image and shows only the luminance value.	
Luminance High (LH)	In this filter, the luminance of the image is increased by applying a gamma correction (Pratt, 2001) to the luminance value. This allows the screeners to see details in dark areas of X-ray images, but as a consequence the visibility of details in light areas of the images is reduced.	
Luminance Low (LL)	As the opposite of the Luminance High filter, the luminance of the image is decreased. Details in light areas of the image become more visible, dark areas lose the details.	
Luminance Negative (LN)	In the Luminance Negative filter, the luminance of the image is inverted. The material value and therefore the hue of each pixel remains the same.	

Continued on Next Page...

Table 4.1: Image Enhancement Filters

Metal Only (MO)	Here, only the metallic parts of the image are shown in color. The organic parts are transformed to light gray with low contrast. The organic parts of the mixed organic/metallic pixels are removed as well, giving them a blue color similar to the all-metallic parts. The motivation for this filter is to allow the screeners to concentrate on the metallic objects perhaps leading to less search time for such objects.	
Metal Stripping (MS)	The Metal Stripping filter removes the metal from the image. Metallic parts are transformed to light gray and from the mixed organic/metallic pixels the metallic part is removed. As some mixed organic-metallic parts originate from metallic objects laying upon organic objects, this removal of metal sometimes shows the complete organic object without potentially distracting metallic parts.	
Organic Only (OO)	The Organic Only filter shows the organic parts of the image in color, while the metallic pixels are set to gray. The mixed organic/metallic pixels are assigned to the metallic or organic parts depending on the proportion of metallic and organic material. The difference to the Metallic Stripping filter is that less of the image remains visible and that the remaining mixed organic/metallic pixels are still green.	
Original (OR)	Original (OR) refers to the unaltered images as produced by the x-ray screening machine without applying any image enhancement filter.	
Organic Stripping (OS)	As the opposite to the Organic Only filter, the metallic parts of the image remain colored and the organic parts are shown in light gray with low contrast. The resulting image is similar to the Metal Only image, except that in this filter the mixed organic/metallic pixels are still green.	

Continued on Next Page...

Table 4.1: Image Enhancement Filters

Super Enhancement (SE)	The Super Enhancement filter adaptively adjusts the contrast of the image. Similar to a Local Histogram Equalization (Gonzalez & Woods, 2002) or an Adaptive Contrast Enhancement (Stark, 2000), the luminance of each pixel is adjusted to the luminance of its surrounding pixels. In the resulting image, each area has a medium average luminance.	

4.2 Experiment 1

Experiment 1 was conducted to evaluate IEFs available in conventional cabin baggage screening (CBS).

4.2.1 Participants

A total of 443 airport security screeners of the CBS at a European airport participated in this study. All had on-the-job experience of at least 6 months. A between-participants design was used to compare the effect of the IEFs on detection performance with each other. To this end, participants were randomly assigned to one of nine experimental groups, one group for each of the nine IEFs specified in Table 4.1. The control group was used for testing detection performance when images were displayed using the Original (OR) image type. The assignment of participants to groups was conducted so that the distribution of gender, age, and days on job were equal across groups. The ten groups showed an equal average of detection performance A', which was calculated using data of a separate test conducted prior to this study. The experimental groups varied in size between 37 (Luminance Negative filter) and 66 screeners (Grayscale filter); the control group consisted of 39 screeners.

The difference in the group sizes is due to missing values (i.e., incomplete tests) for several screeners who originally were assigned to the study.

4.2.2 Method and Procedure

The X-Ray Competency Assessment Test (X-Ray CAT) was used in Experiment 1. This computer-based test contains 256 X-ray images of real passenger carry-on bags. Half of these images contain one prohibited item. The prohibited items have been selected by police experts to be representative for the variety of different threat types. The test contains 32 X-ray images of passenger bags with guns, 32 images with knives, 32 images with improvised explosive devices (IEDs), and 32 images with other prohibited items. For further details on the X-Ray CAT, see Koller and Schwaninger (2006). In order to create the stimuli for Experiment 1, the nine IEFs explained in Table 4.1 above were applied to the X-ray images. The participants' task is to visually inspect the images and to judge whether they are OK (contain no prohibited item) or NOT OK (contain a prohibited item). In this study, images disappeared after 10 seconds. The experiment consisted of two blocks. In block 1, each of the 9 experimental groups was tested with only one IEF and the control group was tested with the Original image (OR). The purpose of block 2 was to confirm that the participant groups are equivalent regarding their X-ray image interpretation competency. In block 2, all participants were tested again using the same bags as in block 1 but images were displayed in the OR format (see Table 4.1).

4.2.3 Results and Discussion

Detection performance was measured using A', a measure derived from hit and false alarm rates (Pollack & Norman, 1964; see Hofer & Schwaninger, 2004, for X-ray image interpretation competency). The hit rate refers to the proportion of all images containing a prohibited item that have been judged as NOT OK. The false alarm rate refers to the proportion of NOT OK judgments for harmless bags. A' scores were calculated for each block separately. Figure 4.1 shows means and standard errors of A' scores of block 1 broken up by IEF and pooled across threat categories,

including the results of the control group (OR). The results in Figure 4.1 suggest that the OR image type results in the best performance, while some IEFs result in substantial impairment of detection performance. Note that due to security reasons, A' scores are not shown in the figures. To estimate effect sizes we employ effect size analysis and interpret the results based on Cohen (1988).

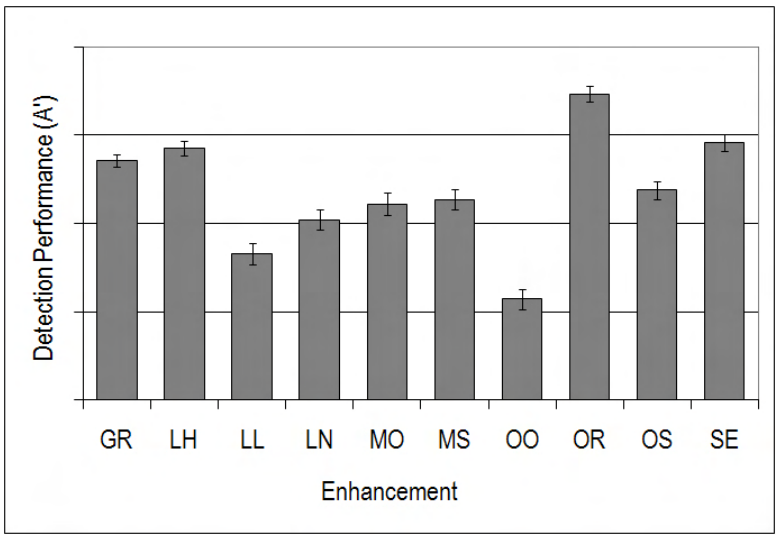


Fig. 4.1. Detection performance Experiment 1, block 1, pooled across threat categories. IEFs were tested between participant groups: GR = Grayscale, LH = Luminance High, LL = Luminance Low, LN = Luminance Negative, MO = Metal Only, MS = Metal Stripping, OO = Organic Only, OS = Organic Stripping, SE = Super Enhancement, OR = Original (control group)

An analysis of variance (ANOVA) with the between-participants factor IEF was carried out on individual A' scores averaged per screener across threat category. There was a main effect of IEF with a large effect size of $\eta^2 = .46, F(9, 433) = 41.67, p < .001$. Figure 4.2 shows means and standard errors of A' scores of block 1 broken up by IEF and threat category. For all four threat categories, the OR image type resulted in the best performance. Again, some IEFs impaired detection performance substantially. Moreover, the results in Figure 4.2 suggest that the effects of IEFs on performance vary between threat categories. These results were confirmed by a separate ANOVA using individual A' scores calculated for each of the four threat categories (guns, knives, IEDs, other prohibited items). The ANOVA with the between-participants factor IEF and the within-participants factor threat category

gave a large main effect of IEF with an effect size of $\eta^2 = .48, F(9, 433) = 43.66, p < .001$. There was also a large main effect of threat type with an effect size of $\eta^2 = .30, F(3, 1299) = 180.84, p < .001$. And there was also a large interaction between threat category and IEF with $\eta^2 = .32, F(27, 1299) = 22.91, p < .001$. The same A' scores were subjected to one-way ANOVAs that were conducted separately for each threat category. There was a large main effect of IEF for all threat categories. For guns, there was an effect size of $\eta^2 = .64, F(9, 433) = 86.09, p < .001$, for IEDs $\eta^2 = .32, F(9, 433) = 22.38, p < .001$, for knives $\eta^2 = .32, F(9, 433) = 23.10, p < .001$, and for other prohibited items $\eta^2 = .43, F(9, 433) = 36.27, p < .001$.

In short, the OR image type resulted in the best performance for all threat categories. Moreover, some IEFs resulted in a substantial impairment which clearly depended on threat category. This interaction would be predicted if one takes into account that color information in X-ray images represents different materials and that different prohibited items vary in their material composition. For example, the Metal Only (MO) filter removes organic material from the X-ray image (see also Table 4.1). Since guns and knives usually consist of metallic material, their pixels in the filtered X-ray image remain largely unaffected when the MO filter is used. However, explosive material of IEDs is organic, thus it is not surprising that the MO filter results in a large impairment of IED detection (see Figure 4.2). A similar explanation applies to the effect of the Organic Stripping (OS) filter. When this filter is applied, all metallic parts of the image remain colored and the organic parts are shown in light gray with low contrast. The resulting image is similar to the MO image, except that for this filter the mixed organic/metallic pixels are still green. Since the Metal Stripping (MS) filter removes metallic information from the image, this IEF results in a substantial impairment of the detection of guns and knives, which usually contain much metal. Because organic explosive material in IEDs remains visible when the MS filter is used, IED detection is not affected substantially. The results in Figure 4.2 also indicate that the MS filter might be a better option than the Organic Only (OO) filter. As explained in Table 4.1, the MS filter includes information about organic material hidden behind metallic parts, whereas the OO filter simply removes these parts from the image. A comparison between the original

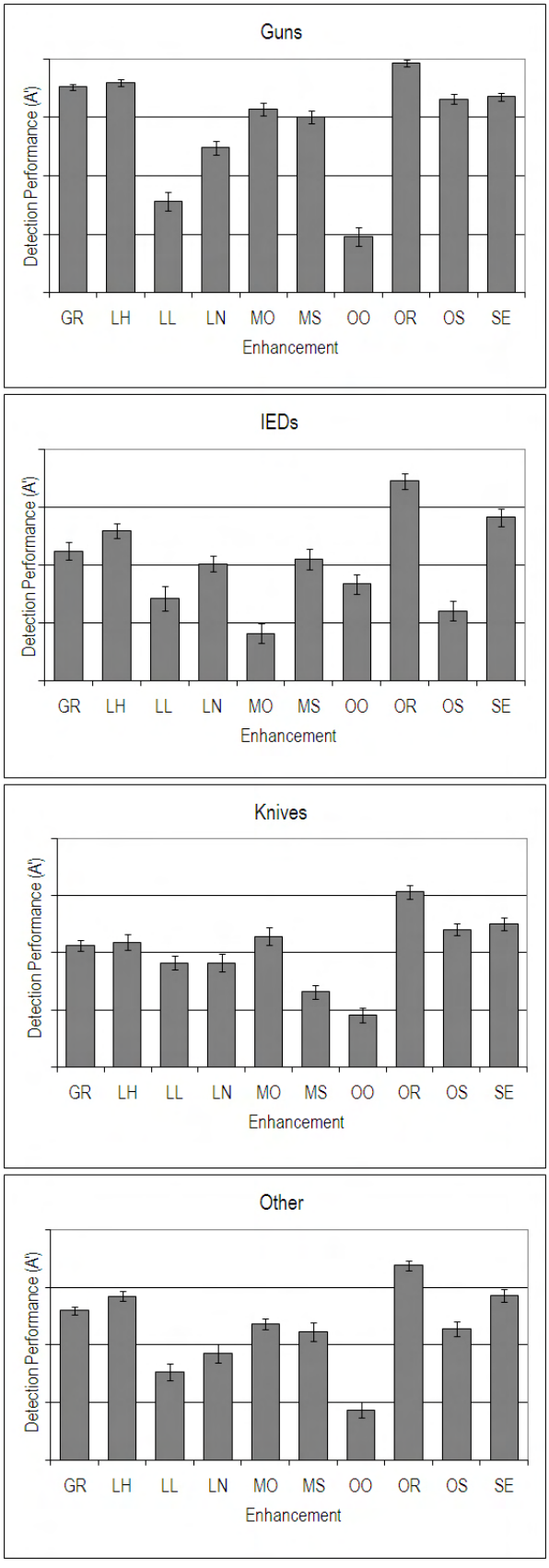


Fig. 4.2. Detection performance in Experiment 1, block 1, broken up by threat category.

image (OR) and the grayscale version gives some indications on the relevance of color information. The removal of the color-coded material information by the Grayscale filter (GR) does impair threat detection, while this effect is less pronounced for the detection of guns. Apparently, the luminance information seems to be more important than the material information. When inserting a threat object into a bag, the part of the bag with the object inside normally becomes darker than its surrounding. This is particularly the case for guns which contain much metallic material. Note however, that the removal of material information can conceal objects with the same luminance but different material than its surrounding. A similar problem appears when using the Super Enhancement (SE) filter. For this IEF, the material information remains the same, but the luminance contrast is slightly reduced which has a negative influence on detection performance. The Luminance High (LH) filter allows better threat detection than the Luminance Low (LL) filter. With the LL filter, most objects inside the bag have a luminance close to black, which generally reduces the differentiation of these objects. When using the Luminance Negative (LN) filter, material and luminance information remain in the image, but the luminance is inverted. The impairment of threat detection when using this IEF shows that screeners perform better with a dark object on a light background than if the luminance is inverted. The results reported so far refer to block 1. As explained in the method section above, all participants conducted the X-Ray CAT again in block 2 using the original image type (OR). This was conducted to confirm post-hoc that the different participant groups are equivalent in terms of their X-ray image interpretation competency. This a prerequisite for the interpretation of the results reported above involving ANOVAs with IEF as between-participants factor. Separate ANOVAs of the data from block 2 confirmed that the 9 experimental groups and the control group were equivalent. Individual A' scores were calculated for each screener based on all trials of block 2. These data were subjected to a one-way ANOVA with participant group as between-participants factor. All groups were equivalent, since there was no effect of group, $\eta^2 = .02, F(9, 433) = 1.08, p = .38$. Individual A' scores were calculated also for each threat category separately and this data were then analysed using an ANOVA with participant group as between-participants fac-

tor and threat category as within-participants factor. Again, the results show that the participant groups were equivalent in terms of their X-ray image interpretation competency, since there was no main effect of participant group, and no interaction between participant group and threat category, $\eta^2 = .02, F(9, 433) = 1.04, p = .41$, and $\eta^2 = .02, F(27, 433) = 0.86, p = .63$, respectively.

4.3 Experiment 2

In hold baggage screening (HBS) X-ray images feature slightly different colors. Figure 4.3 shows examples of the stimuli used in Experiment 2. As explained in the introduction, screeners mainly search for IEDs, as other threat objects like for example knives do not pose a threat to the aircraft and passengers when placed in hold baggage. HBS screeners are often also more experienced screeners as it was the case in this participant sample. The main aims of Experiment 2 were to examine whether similar results are found in HBS regarding the effect of IEFs despite the operational and training differences between HBS and CBS.

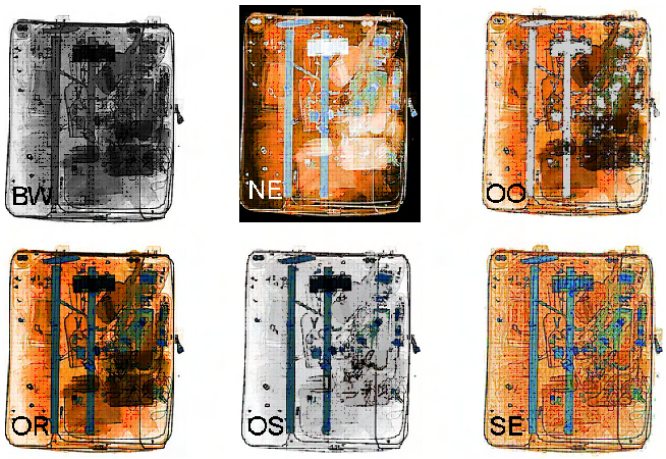


Fig. 4.3. IEFs for HBS as used in Experiment 2. From top left to bottom right: GR, LN, OO, OR, OS, SE (see Table 4.1).

4.3.1 Participants

Data of 83 aviation security screeners of the HBS of the same European airport was analyzed. As in Experiment 1, a between-participants design was used to compare

the effect of the IEFs. Due to the smaller sample size only 5 IEFs and the OR image could be tested. The 83 HBS screeners were randomly assigned to one of five experimental groups (GR, LN, OO, OS, SE filters) or the control group (OR filter). The assignment of participants to groups was conducted so that the distribution of gender, age, and days on job was equal across groups. The six groups showed an equal average of detection performance A', which was calculated using data of a separate test conducted prior to this study. The number of screeners in each experimental group were between 10 (GR) and 17 (OO); the control group (OR) consisted of 15 screeners. As in Experiment 1, the difference in the group sizes is due to missing values (i.e., incomplete tests) for several screeners.

4.3.2 Method and Procedure

The Bomb Detection Test (BDT) was used in this study. This computer-based test contains 200 X-ray images of real hold baggage, whereas 100 images contain an IED. The IEDs were created by police experts. Participants were instructed to decide for each X-ray image whether it is OK (does not contain an IED) or NOT OK (contains an IED). Images disappeared after 10 seconds. As in Experiment 1, there were two blocks. In block 1, each of the 5 experimental groups was tested with their respective IEF. In block 2 all participants were then tested again using the same images but using the Original (OR) image function. The control group conducted the test twice using the OR image type in block 1 and block 2. As in Experiment 1, the purpose of block 2 was to confirm the comparability of the groups post hoc.

4.3.3 Results and Discussion

Analyses were similar to Experiment 1 but there was only one threat category, namely IEDs. Figure 4.4 shows means and standard errors of A' scores broken up by image enhancement function. As mentioned above, A' scores are not shown in the figure for security reasons. Effect sizes are calculated using effect size analysis and they are interpreted based on Cohen (1988). A one-way ANOVA with IEF as between-participants factor revealed a large main effect of IEF with an effect size

of $\eta^2 = .26, F(5, 77) = 5.29, p < .001$. As in Experiment 1, the original image (OR) resulted in the best performance. Consistent with the results found in Experiment 1, we found in Experiment 2 that the Organic Stripping (OS) and Luminance Negative (LN) functions resulted in a substantial impairment of detection performance for IEDs. All participants conducted the test again in block 2 using the original image type (OR). The aim was to confirm post-hoc that the different participant groups are equivalent in terms of their X-ray image interpretation competency. To this end, individual A' scores from block 2 were subjected to a one-way ANOVA with participant group as between-participants factor. There was no main effect of group, $\eta^2 = .05, F(5, 77) = 0.75, p = .59$, confirming that the six groups are equivalent regarding their X-ray image interpretation competency.

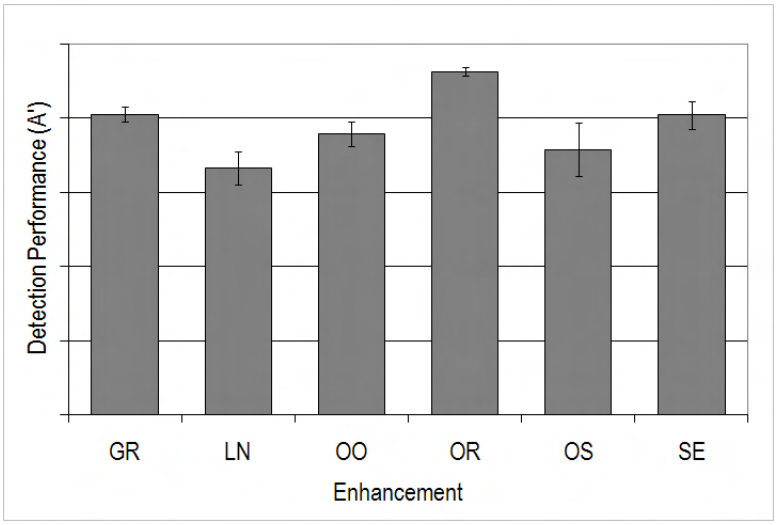


Fig. 4.4. Detection performance Experiment 2, block 1. GR = Grayscale, LN = Luminance Negative, OO = Organic Only, OR = Original, OS = Organic Stripping, SE = Super Enhancement.

4.4 General Discussion

The aim of this study was to investigate the effect of image enhancement functions (IEFs) on X-ray detection performance of airport security screeners. Experiment 1 was conducted with cabin baggage and Experiment 2 with hold baggage. In both experiments the original image (OR) resulted in the best performance. One inter-

pretation could be that for this manufacturer the default image is indeed the best image. However, since the OR image is the default image on the tested X-ray machine and since screeners received more training with OR images, further research is needed to clarify whether the benefit of the OR image type is due to expertise and training or whether it truly reflects better image quality. In both experiments, it was also found that some IEFs resulted in substantial impairments of detection performance. This general result is consistent with previous reports (Klock, 2005; Schwaninger, 2005c). The IEF effects are dependent on threat category; most likely due to differences in material properties of the different threat categories. For example, guns contain more metal than IEDs. Removing metallic content (MS function) therefore results in a larger impairment of detection performance for guns than for IEDs. The main conclusions of this study are that user testing is crucial before implementing such filters into a system. Moreover, training when and how to use each of the filters is crucial to make effective use of them. We are conducting a set of additional experiments to further investigate the value of IEFs. For example, it could be that although on average certain IEFs impair detection, they could still be useful for detecting certain threat objects under certain conditions. Moreover, we are currently looking at CBT data where screeners have the possibility to choose a filter and to switch between filters. This allows investigating whether perhaps a certain combination and sequence of IEFs is useful for certain threat types and images. In addition, we are trying to clarify, whether IEFs actually do not improve detection of prohibited items or if however, when used according to individual preferences and to specific features of the image, they can improve the ability to locate targets. Finally, we also have implemented IEFs in a CBT system (X-Ray Tutor) to investigate potentially supporting effects that only can become manifest through training and familiarization.

Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). Many thanks to Zurich

State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

5

The Role of Consideration Information for X-ray Image Interpretation in Aviation Security

5.1 Introduction

The task of aviation security screening officers at an airport comprises many different components. The main task is to prevent passengers from bringing prohibited objects onto the plane. The prohibited objects are defined by Doc 30 of the European Civil Aviation Conference (ECAC) and include - among others - for example guns, knives, and improvised explosive devices. These objects are prohibited because they could pose a danger to the safety of passengers during a plane travel. In order to carry out this task efficiently all passenger bags have to go through an X-ray screening machine. In this way, screening officers do not have to manually open and search every single suitcase. Instead, they search the X-ray image of the passenger bag and decide based on this information if a bag is clear to be carried onto the plane or if it contains a suspicious object. If this is the case, or if the interpretation of the X-ray image is not possible due to high density of the content, the bag has to be opened and searched manually and appropriate measures have to be taken, respectively. The aim of the X-ray screening process is to detect all prohibited objects with minimum number of bags to be opened. In order to achieve this goal, the screening officers have to be able to discriminate between threat objects and harmless everyday objects when they see them on the X-ray screen. At the moment, screening officers at many airports are supposed to conduct individually adaptive computer-based training (for example with X-Ray Tutor (XRT), Schwaninger, 2004) in order to enhance their skills in detecting threat objects and in discriminating them from harmless

everyday objects. Detection of threat objects and object recognition in general is dependent on several factors. On one hand, screening officers have to know which objects are prohibited and what they look like in X-ray images. Object shapes that are not similar to ones that are stored in visual memory are difficult to recognize (e.g., Graf et al., 2002; Schwaninger, 2004, 2005a). This is the knowledge-based factor. Detection of (threat) objects is also depending on image-based factors (Schwaninger et al., 2005). Image-based factors refer to attributes of an object that change the difficulty of its detection or recognition if they are varied. Difficulty to recognize an object can be increased if it is seen in an unusual angle, that is, if the viewpoint is varied. Various studies on object recognition found orientation effects not only for novel objects (e.g., Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Tarr & Pinker, 1989) but also for known, familiar objects (e.g., Hayward & Tarr, 1997; Lawson & Humphreys, 1996, 1998; J. E. Murray, 1997, 1999; Newell & Findlay, 1997; Palmer et al., 1981). The view of an object in which it is easiest to recognize is often referred to as canonical view (Palmer et al., 1981). Furthermore, an object becomes more difficult to recognize if it is superimposed by other objects. With increasing superposition characteristic features of an object might get lost which are essential to recognize this object as what it is. Several studies found reduced object recognition for incomplete pictures (e.g., Biederman, 1987; F. S. Murray & Szymczyk, 1978). Another factor influencing the recognizability of an object is its surrounding. In the case of passenger bags it is defined as bag complexity which depends on the type and number of other objects in the bag. Too much information distracts attention and impedes detection and recognition of objects. The computer-based training system XRT incorporates these factors in its individually adaptive algorithm. Screening officers see X-ray images of passenger bags and have to judge if a bag is OK (contains no threat object) or if it is NOT OK (contains threat object). Every screening officer starts at the first level and reaches the next level by accomplishing a certain detection performance. This means, enough threat objects have to be recognized and detected, but enough clear bag images have to be judged as OK as well (as evidence of discrimination ability). In each level the difficulty of these images increases. First, the viewpoint is exacerbated. Then, with increasing levels, superposition and

bag complexity are also exacerbated. In the highest level screening officers see very complex bags containing - if - highly rotated threat objects which are superimposed by other objects. For more details on XRT and the individually adaptive algorithm please refer to Schwaninger (2004) and Schwaninger, Michel, and Bolfing (2007). A study by Koller et al. (2008) has shown that there is a large increase of detection performance when screening officers are trained with XRT. In this study another approach is examined to support the screening officers in their decision about OK or NOT OK while they search an X-ray image of a passenger bag. The aim of this study is to investigate whether displaying consideration information can help security screening officers to identify prohibited items in X-ray images of passenger bags, and to make a decision if a detected object actually is a threat object or not. The use of consideration information serves the intention to influence the screening officers to make fewer mistakes regarding the detection and identification of threat objects. Consideration information refers to simultaneously presented X-ray image patches which contain visually and conceptually similar prohibited items as well as X-ray image patches which are visually similar but do not contain a prohibited item. The role of consideration information for increasing detection performance and efficiency is investigated with novices, intermediate and expert screeners and for the time being focuses on the detection of improvised explosive devices (IEDs). The classification into novices, intermediate, and expert screeners is done under the assumption that consideration information may have a different effect depending on expertise. The concept of consideration information is derived from the mainly marketing research related concept of the consideration set. The consideration set is a subdivision of a consumer's awareness set, which constitutes all brands available on the market of which the consumer is aware of. Those brands out of the awareness set that the consumer would consider purchasing is called the consideration set. The purchase decision is restricted to brands belonging to the consideration set (e.g., Kotler, 1988; J. H. Roberts & Lattin, 1991, 1997). There is a substantial difference between the two concepts in that the consideration set is a model to describe consumers' (human subjects') layered decision process whereas the consideration information represents the context to influence human subjects' decision. In fact, the concept of consider-

ation information has its root in consumer behavior research where context effects are used to influence human subjects' choices (e.g., Hamilton, 2003). There is little literature available on consideration information as help in visual search and decision making. Among these, Bernstein and Li (2005) investigated the role of consideration information on a digit recognition task and found evidence that the consideration information did improve the participants' performance.

5.2 Method

5.2.1 Participants

The study was conducted at a large European airport with 294 screening officers participating voluntarily. After eliminating incomplete data sets, data of 198 was used for analysis. The screening officers were divided into three groups: novices, intermediate, and expert screeners. Since there is no explicit definition of novice, intermediate, and expert screener, this assignment was done twice. Once based on the duration of their employment and once based on their performance in the certification test last year. This means, in the first approach, experts are those screening officers with the longest working experience. In the second approach, experts are those screening officers with the best X-ray image interpretation competency. Working experience was assessed based on the duration of employment and calculated in days on job (DOJ). Thereby, novices were defined as persons working less than 1.5 years (i.e., 547.5 days) in this job. Screening officers working between 1.5 and 5 years (i.e., between 547.5 and 1825 days) in the job were classified as intermediate screeners, and experts are those screening officers who have been working more than 5 years (i.e., more than 1825 days) in the job. These definitions are chosen arbitrarily and in a way that the distribution of the screening officers yields comparable group sizes. See Table 5.1 for details. Table 5.1 also displays the average and standard deviation of days on job for the three groups as well as for the three status groups defined by the performance approach. The other approach defines the expertise status on performance instead of time on job. Therefore, results of an X-ray image test (X-Ray CAT, see Koller

& Schwaninger, 2006) provide the basis for this assessment. The X-Ray CAT was solved by every screening officer in the course of their yearly certification in 2007. Performance was measured using A'. Mean A' in X-Ray CAT for the experimental group was $m = 0.910$, SD = 0.033. Based on these values expert status is defined as a minimum A' $\geq (m+0.5SD)$, that is, 0.926. Novice status is defined as a maximum A' $\leq (m - 0.5SD)$, that is, 0.893. Screening officers that achieved an A' between 0.893 and 0.926 in the X-Ray CAT are classified as intermediate. See Table 5.1 for details. Table 5.1 also displays mean A' values from the X-Ray CAT for each group.

Table 5.1. Figures of the expertise groups, defined by days on job (DOJ, top half) and by test performance (CAT, lower half). SD = Standard deviation.

Status DOJ	Definition	n	Mean DOJ	SD of DOJ	Mean A'	SD of A'
Expert	DOJ > 1825	69	4592.9	2140.4	0.909	0.028
Intermediate	DOJ 547.5 - 1825	63	966.0	332.4	0.915	0.038
Novice	DOJ < 547.5	66	357.4	140.8	0.903	0.030
Status CAT	Definition	n	Mean DOJ	SD of DOJ	Mean A'	SD of A'
Expert	A' > 0.926	60	2187.0	2213.0	0.942	0.012
Intermediate	A' 0.893 - 0.926	68	2297.6	2440.7	0.910	0.010
Novice	A' < 0.893	70	1627.1	2163.9	0.868	0.024

5.2.2 Materials and Procedure

Stimuli were created from Smiths-Heimann Hi-Scan 6040i colour X-ray images of improvised explosive devices (IEDs) and passenger bags (Figure 5.1 displays an example of the stimuli). Ten prototypical IEDs were chosen and manually built into four bag images each, controlling for the superposition. This was done with a tool which applies a special algorithm when combining two X-ray colour images.

Using the following formula, this tool calculates the superposition of the built-in prohibited item, that is, how much the pixels of the IED in this case are superimposed by pixels of the bag:

$$SP = \frac{\sqrt{\sum[I_{SN}(x,y) - I_N(x,y)]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); ObjectSize: Number of pixels of the prohibited item where R, G, and B are < 253

Superposition is kept constant for each IED in the four different bags. In the experiment the examinees see all 40 bags containing an IED (threat images) and additionally the same 40 bags but not containing an IED (non-threat images). The consideration information image patches are created by combining principle component analysis (PCA) (Jolliffe, 2002) and manual IED insertion.

Fig. 5.1. Left: main IED. Middle and right: reference IEDs as used in consideration information.

Specifically, first, for each threat image the image patch containing the IED was cut out (same size for all items). Formally, we name them threat image patches. Then, the threat image patches are used to search for **visually similar image patches** through a large set of non-threat passenger bags (empty bags). In principal, similarities among the images are measured by projecting the images into a constructed feature space and defining a distance over the

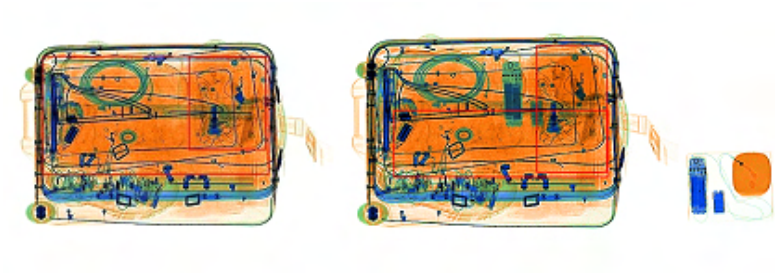


Fig. 5.2. Example images. Left: harmless bag (non-threat image), right: same bag with an improvised explosive device (IED) at the top right corner (threat image). The IED is shown also separately at the bottom right.

feature space. There are many ways to construct features (e.g., spreading from low level feature detection of edge, corner, blob, ridge to more complex features related to texture, shape or motion) from the images depending on the desired target (e.g., face detection) (Morris, 2004). Since we are interested in the general visual similarities among images, the general dimensional reduction techniques (e.g., PCA, semidefinite embedding, multifactor dimensionality reduction, Isomap, Kernel PCA, etc) become suitable tools to construct features. It is unclear which dimensional reduction technique should be chosen due to the absence of the quantitative visual similarity measurement criterion, and it is not a topic of this paper, either. To focus on our interests, we optionally pick up the simple PCA as the feature construction tool and adopt a Euclidean distance to define the similarities among X-ray images. Formally, we denote the threat image patches as X_i, and the non-threat passenger bag set as X_E. Then the following procedure is used to compute the similarity between a threat image patch $x_i \in X_i$ and every non-threat passenger bag $x_e \in X_E$.

PCA-Based x-ray Image Similarity Computation Procedure

Input: X_i, X_E

1. Combine X_i and X_E together and denotes the combined image set as $X = [X_i, X_E]$, where each column of X represents an image.
2. Compute the covariance matrix of X as $COV = (X - \overline{X})(X - \overline{X}^T$. Note, $\overline{X}$ denotes the mean matrix of X, where each column of $\overline{X}$ has the identical value, which equal to the average intensity of corresponding image in X.
3. Compute 30 maximum eigenvalues $\lambda = [\lambda_1, ..., \lambda_{30}]$ and corresponding eigenvectors $Q = [q_1, ..., q_{30}]$ of COV.
4. Project X into feature space Q by a dot product $Q^T(X - \overline{X})$. Here we denote the projection of X as X_{PCA}, where each column of X_{PCA} represents a 30-PCA transformation of the corresponding original image.
5. For each $x_{i,PCA} \in X_{i,PCA}$,
   a) loop through $X_{E,PCA}$, for each $x_{e,PCA} \in X_{E,PCA}$, compute Euclidean distance between $x_{i,PCA}$ and $x_{e,PCA}$ as $d_{i,e} = (x_{i,PCA} - x_{e,PCA})^T(x_{i,PCA} - x_{e,PCA})$

b) select 8 different $x_{e,PCA} \in X_{E,PCA}$, which have the minimum Euclidean distances from $x_{i,PCA}$ and store the corresponding images as 8 most similar non-threat passenger bags for a given threat image patch x_i

With above technique, for each threat image patch 8 similar looking patches from the empty bag selection were extracted. These 8 image patches were distributed randomly to the appropriate threat and non-threat image pair (4 for the threat image and 4 for the non-threat image). In the experiment, each image is presented with 4 consideration information image patches, whereof 2 are images containing a reference IED and 2 are empty images. For each of the 10 main IEDs 2 reference IEDs were selected which are similar concerning the visual appearance and the structure of the IED. An example of a main IED with the two reference IEDs is depicted in Figure 5.2. Both reference IEDs are built into the consideration information image patches and presented with the main IED, accompanied by the two empty consideration information image patches. Figure 5.3 shows the user interface of the experiment. The non-threat images are presented with 4 consideration information image patches as well, 2 of them containing the reference IEDs corresponding to the belonging threat image.

The images are divided into two blocks (Block 1 and Block 2) and presented randomly, where each threat and non-threat image pair is split up in the two blocks. Screening officers had the possibility to activate the experiment during a training session by clicking the corresponding button on the screen. They were randomly assigned to one of three groups with different experimental designs (Consideration Information, Feedback, and Control condition) and two blocks within each experimental group. Screening officers in Block 1 started with images of Block 1 and then saw the images of Block 2 and vice versa for screening officers in Block 2. The block design is applied to avoid a threat image to be followed immediately by the corresponding non-threat image. These data will not be analyzed within this study. The task is the same for all three experimental conditions: the main image has to be judged whether it contains an IED (click Nicht OK) or if no IED is enclosed in the bag (click OK) (see Figure 5.3 for the user interface). Furthermore, the screening officers had to rate their confidence of the judgment on a sliding bar (not analyzed

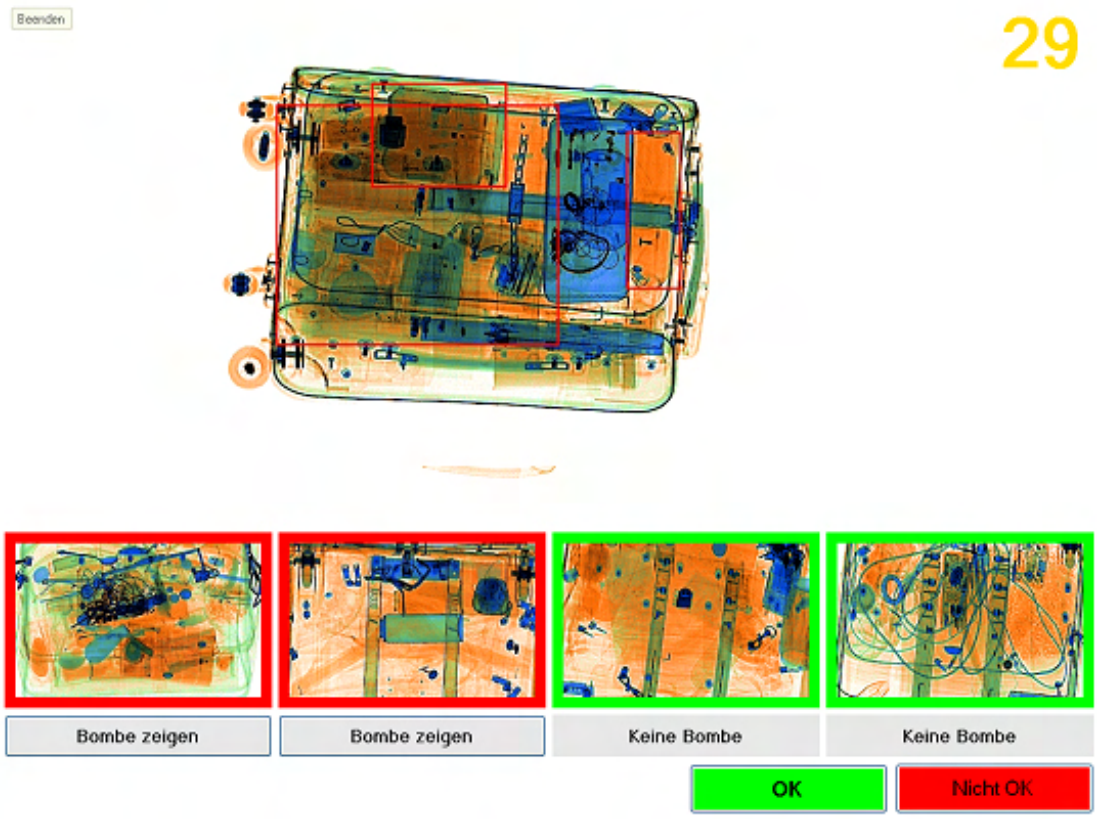


Fig. 5.3. Screenshot of the Consideration Information Test Platform. The big bag image has to be judged if it is "OK" (contains no IED) or "Nicht OK" (NOT OK, contains IED). The four little image patches on the bottom are consideration information. The images with the red border contain a reference IED, the images with the green border contain no IED. By clicking on the button "Bombe zeigen" (show bomb) the IED could be highlighted with a flicker. A timer on the top right indicates the seconds the main image remains on the screen.

in this paper). This confidence rating was only activated after the response OK or Nicht OK (NOT OK) was given. The main image was displayed for 30 seconds but responses could also be given when the main image was removed. In the Consideration Information condition (n = 68), the consideration information is presented simultaneously to the main image, which has to be judged. By clicking on "Bombe zeigen" the IED in the corresponding consideration information image patch can be highlighted through a flicker (see Figure 5.3). In the Feedback condition (n = 66), the consideration information is presented only after the response (OK or Nicht OK) is given. This is to rule out any effect on the results of presenting information. In the Control condition (n = 64) no consideration information was presented at all.

5.3 Results

Hits, misses, false alarms and correct rejections were reported for each screener's response to each single trial. Hits are defined as all correctly identified bombs (i.e., the response NOT OK to a threat image). Misses are missed bombs (i.e., the response OK to threat images). False alarms are NOT OK responses to non-threat images and correct rejections are OK responses to non-threat images. Individual data were averaged across all images in order to eliminate an item-specific factor. Hit rate and false alarm rate were calculated. The hit rate is the proportion of all threat images that were correctly judged as containing a bomb. The false alarm rate is the proportion of all non-threat images that were incorrectly judged as containing a bomb. Using these two parameters and applying signal detection theory the detection performance measure d' was calculated: $d' =$ zHit - zFalse, where zHit and zFalse are the standardized hit and false alarm rates. Furthermore, reaction times were analyzed as well, that is, the time during which the main image is presented on the screen until a response was given. The analyses regarding status group (novices, intermediate and expert screeners) were made twice, once with the status classification based on working experience (Status DOJ) and once with the status classification based on X-ray image interpretation competency (i.e., certification results, Status CAT). Table 5.2 gives an overview on the number of screeners per group, once according to the classification by working experience, once according to the classification by performance.

Table 5.2. Number of screeners for the different expertise groups in the three experimental conditions Consideration Information, Feedback, Control

		Status DOJ	Status CAT
Consideration Information	Expert	28	17
	Intermediate	22	31
	Novice	18	20
Feedback	Expert	24	22
	Intermediate	21	21
	Novice	21	23
Control	Expert	17	21
	Intermediate	20	16
	Novice	27	27

5.3.1 Detection Performance

Figure 5.4 shows the average detection performance *d'* for the three experimental groups, independent of the status (*Note*: All *d'* values are multiplied with an arbitrary constant for security reasons). A univariate Analysis of Variance (ANOVA) with the between-participants factor experimental condition (Consideration Information, Feedback, and Control) confirmed the visual impression that there is no significant difference in detection performance between the three experimental groups ($p = .799$). So in general it can be said that the presentation of consideration information does not help to improve detection performance, in this case of IEDs. Looking

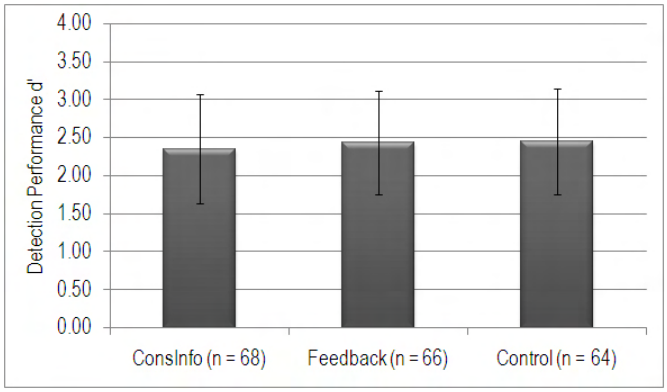


Fig. 5.4. Detection performance for the three experimental conditions: Consideration Information, Feedback, and Control condition. Thin bars are standard deviations.

at the results broken up by DOJ status group (Experts, Intermediate, Novices, see Figure 5.5), performance differences can be noted between the conditions. Indeed it seems that consideration information has a different effect depending on the status group. However, ANOVAs with the between-participants factor experimental condition for each of the status groups revealed no statistically significant differences between the experimental conditions (Experts: $p = .118$; Intermediate; $p = .760$; Novices: $p = .263$). This means, again, that detection performance is not affected by the experimental condition. Still, a tendency can be observed that novices can profit from consideration information, whereas for intermediates there is no effect and for experts it actually impairs detection performance.

If broken up by CAT status group (see Figure 5.6), the pattern seems to be different. However, here as well the ANOVAs with the between-participants factor experimen-

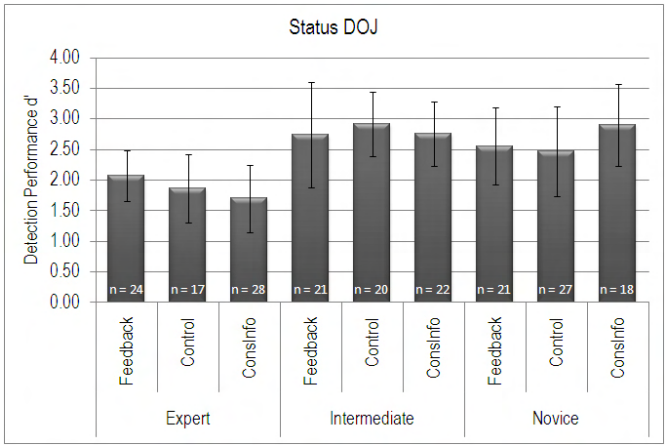


Fig. 5.5. Detection performance broken up by DOJ status group and for the experimental conditions separately. Thin bars are standard deviations.

tal condition for each of the status groups revealed no statistically significant differences between the experimental conditions (Experts: $p = .779$; Intermediate: $p = .916$; Novices: $p = .371$). This is another indication that the detection performance is not affected by the experimental condition and that consideration information in particular did not improve detection performance. Figures 5.7 and 5.8 show scatter

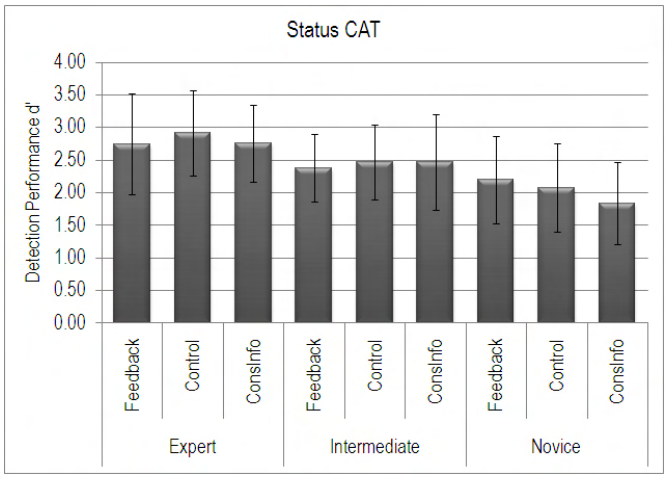


Fig. 5.6. Detection performance broken up by CAT status group and for the experimental conditions separately. Thin bars are standard deviations.

plots of the data from the consideration information condition group only. Figure 5.7 depicts the detection performance for all screeners of the consideration information condition group in relation to their working hours (i.e., working experience). There is a strong negative correlation ($r = -.682$, $p < .01$) between detection performance in this study and days on job. This means, the shorter the working experience, the higher the performance in the consideration information condition. It seems as if inexperienced screeners can profit more from consideration information. However,

it has to be considered that we have no baseline measurement for this group. The correlation between days on job and detection performance for the control group is r = -.289 with $p < .05$ (see Figure 5.9). Figure 5.8 displays the detection per-

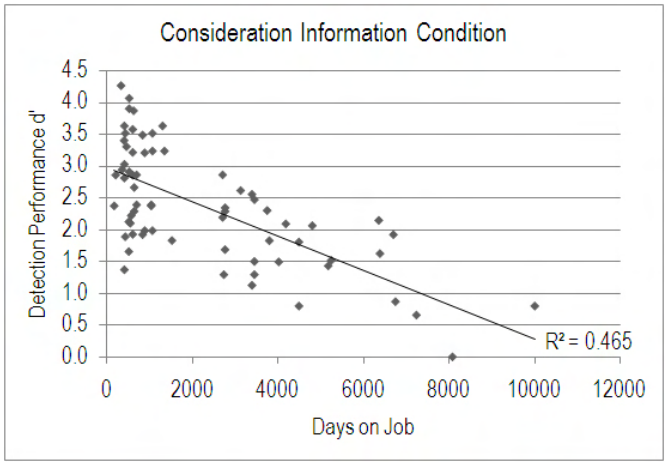


Fig. 5.7. Detection performance *d'* in relation to the days on job for each screener in the consideration information condition group. Each dot represents one screener.

formance in relation to the test performance in X-Ray CAT during certification for each screener in the consideration information condition group. Here, the trend is in the opposite direction. There is a significant correlation ($r = .374$, $p < .01$) between detection performance in this study and test performance in X-Ray CAT, meaning that experts defined as screeners with high X-ray image interpretation competency achieved higher scores in this study. This leads to the assumption that performance-based experts can profit more from consideration information. However, again we have no baseline measurement for these groups. The correlation between test performance in X-Ray CAT and detection performance for the control group is $r = .405$ with $p < .01$ (see Figure 5.10).

5.3.2 Reaction Time

The same analyses as for detection performance have been made for reaction times. Figure 5.11 shows the average reaction times per experimental condition, independent of status. Here, the data suggests that the group with consideration information on average needed most time to respond to the main image, compared to the group receiving feedback and to the control group. In average, the control

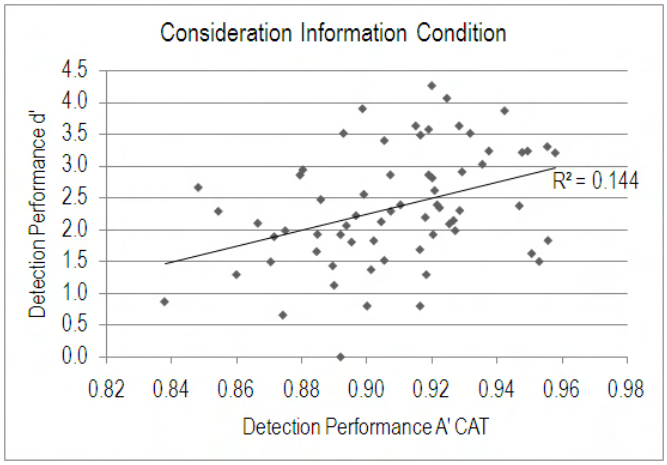


Fig. 5.8. Detection performance *d'* in relation to the detection performance A' in the X-Ray CAT, taken during certification 2007. Each dot represents a screener in the consideration information condition group.

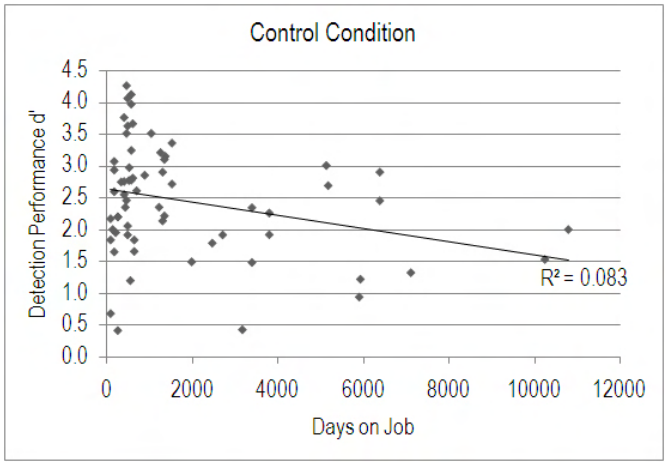


Fig. 5.9. Detection performance *d'* in relation to days on job for the control condition group. Each dot represents one screener.

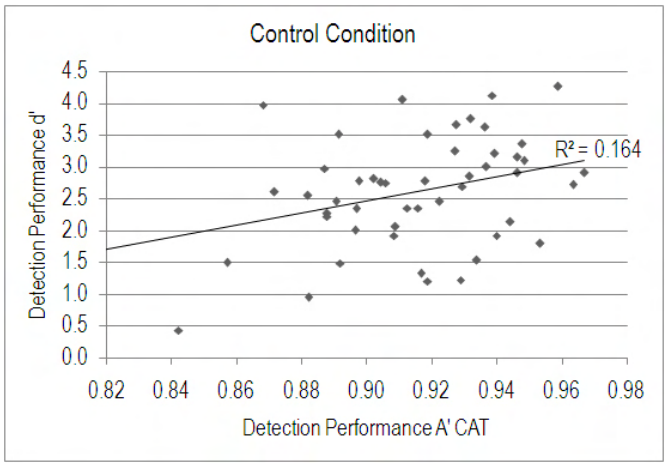


Fig. 5.10. Detection performance *d'* in relation to the test performance in the CAT for the control condition group. Each dot represents one screener.

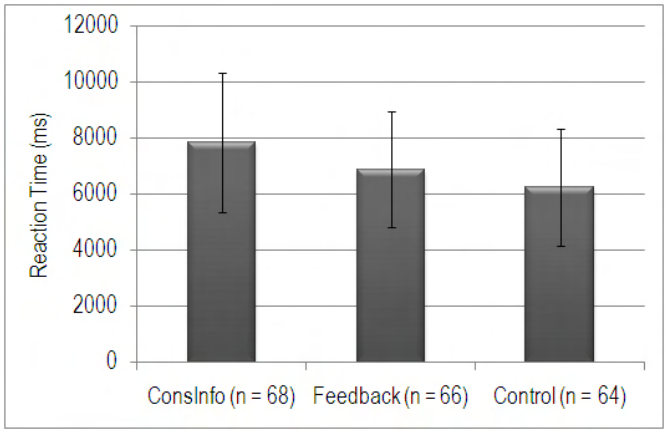


Fig. 5.11. Averaged reaction time in seconds for the three experimental conditions: Consideration Information, Feedback, and Control. Thin bars are standard deviations.

group was fastest to respond to the main image. The ANOVA with the between-participants factor experimental condition confirms a significant main effect with medium effect size ($\eta^2 = .083, F(2, 195) = 8.824, p < .001$). All effect sizes are interpreted according to the conventions by Cohen (1988). Separate pairwise comparisons relate the difference to the consideration information condition, where a significantly higher reaction time was generated than in the feedback condition ($p < .05$) and therefore also than in the control condition ($p < .001$). The difference between the feedback and the control condition is not significant ($p = .096$). Again, analyses were made broken up by status group. Figure 5.12 shows mean reaction time broken up by DOJ status group. All groups needed more time for the consideration information condition than in the feedback or control condition. ANOVAs with the between-participants factor experimental condition for each status group revealed a just about significant main effect of experimental condition on reaction time for the experts ($\eta^2 = .095, F(2, 66) = 3.467, p < .05$) and novices ($\eta^2 = .101, F(2, 63) = 3.534, p < .05$), with medium effect sizes as well. For intermediate screeners the difference was not significantly different between the experimental conditions. Although there was a significant main effect, separate pairwise comparisons between the conditions revealed no significant difference for the expert group. For the novices only the difference between the consideration information and the control condition was significantly different ($p < .05$).

If broken up by CAT status (see Figure 5.13), the pattern is similar. ANOVAs with the between-participants factor experimental condition for each status group re-

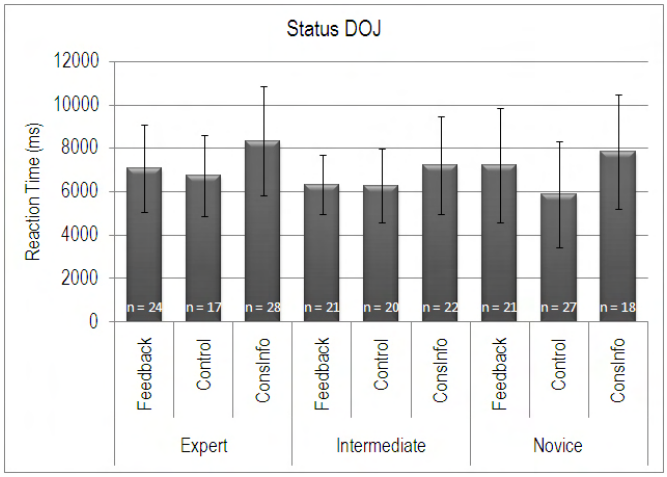


Fig. 5.12. Average reaction time broken up by DOJ status group and for each experimental condition separately. Thin lines are standard deviations.

vealed a large and significant main effect only for the intermediate screeners group ($\eta^2 = .245, F(2, 65) = 10.564, p < .001$). Separate pairwise comparisons of the conditions for the intermediate group showed a significant difference between consideration information and feedback condition ($p < .05$) and between consideration information and control condition ($p < .001$), but not between the feedback and control condition. This means, if status group is defined according to X-ray image interpretation competency, screeners with medium performance need significantly more time for responding to the main image in the consideration information condition. Other than that, there is only the tendency in the same direction for the novices, but no other significant effects of the experimental condition on reaction times.

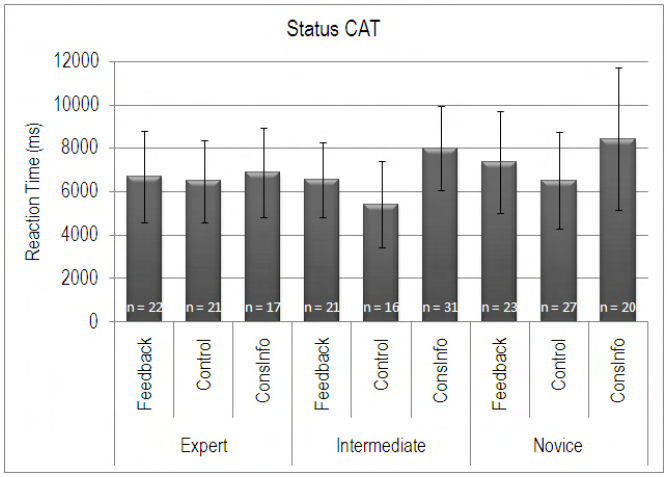


Fig. 5.13. Average reaction time broken up by CAT status group and for each experimental condition separately. Thin bars are standard deviations.

It is also interesting to have a look at the behavior of the screeners during the test regarding consideration information. In the consideration information condition and the feedback condition, where the consideration information was displayed as feedback after the response, screeners had the possibility to highlight the IEDs within the two threat image patches of the consideration information by clicking on the button (see Figure 5.3 for the user interface). Figure 5.14 shows the number of clicks on the two buttons for the consideration information group and the feedback group, averaged over the status groups. It can be seen that the number of clicks on average is considerably higher for the feedback group. However, as could be seen in Figure 5.4, there was no benefit of consideration information or feedback whatsoever for the detection performance.

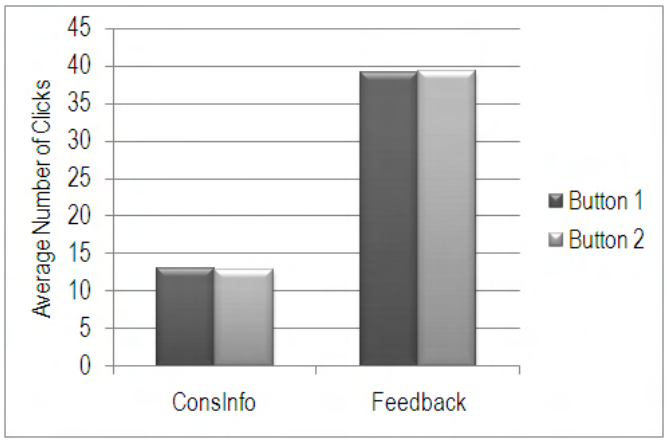


Fig. 5.14. Number of clicks on the two buttons (Button 1 and Button 2) used to highlight the IED within the threat image patch of the consideration information for the Consideration Information group and the Feedback group.

If split up by DOJ status, there is an interesting fact to be observed (see Figure 5.15). In the feedback condition, novices highlight the IEDs in the consideration information far more often than intermediate or expert screeners. However, this difference is not significant, neither between the novices and the intermediate screeners ($p = .125$) nor between the novices and the experts ($p = .075$). This may be due to the large standard deviations which are not depicted in the graph (see Table 5.3 for details). In all cases (except for the intermediate screeners in the feedback condition), standard deviations are larger than the mean values which indicates substantial differences in the use of the buttons between the screeners.

Table 5.3. Average use of Button 1 and Button 2 for the three DOJ expertise groups in the two experimental conditions Consideration Information and Feedback. SD = Standard deviation.

	Status DOJ	Button 1	SD Button 1	Button 2	SD Button 2
Consideration Information	Expert	15.107	25.053	14.429	23.050
	Intermediate	13.273	15.144	13.545	15.121
	Novice	9.722	11.208	9.333	9.677
Feedback	Expert	28.792	39.208	27.333	36.151
	Intermediate	30.238	28.117	29.667	28.472
	Novice	59.667	68.007	62.619	77.486

A different pattern can be seen in the consideration information condition. There is no significant difference between the three status groups regarding use of the buttons. Here as well there are large differences between the screeners regarding the use of the buttons (see Table 5.3 for standard deviations).

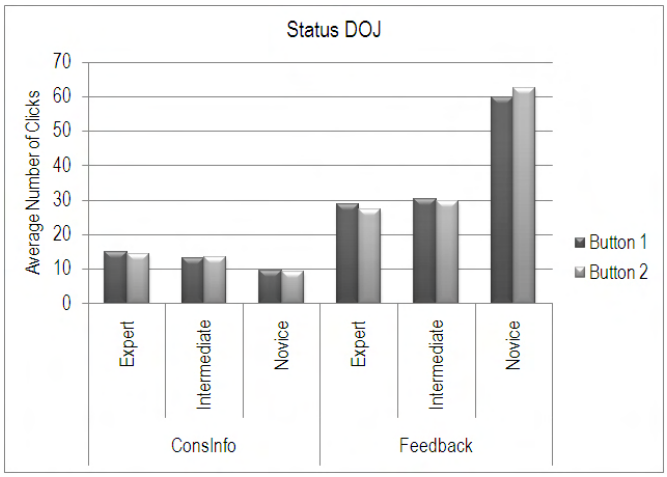


Fig. 5.15. Number of clicks on the two buttons used to highlight the IED within consideration information broken up by DOJ status group.

If broken up by CAT status (see Figure 5.16), the pattern is again similar to the analysis based on working experience. Here, the novices - defined by low X-ray image interpretation competency - are using the highlighting buttons more frequently in both experimental groups. However, the differences between this group and the intermediate and expert screeners are smaller and in no case significant. In general, the tendency shows that, the higher the X-ray image interpretation competency, the fewer the use of the highlighting buttons.

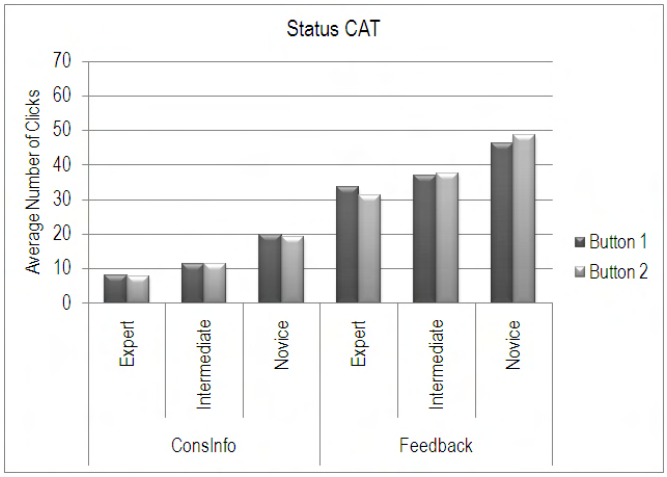


Fig. 5.16. Number of clicks on the two buttons to highlight the IEDs in consideration information broken up by CAT status group for the Consideration Information and the Feedback condition.

5.4 Discussion

The aim of this study was to investigate the concept of consideration information within aviation security screening. Aviation security screening officers have to search every passenger bag for threat objects which are prohibited from being carried onto an airplane. This important task is being executed by means of X-ray screening. That means, screening officers have to judge if a bag is harmless or not by interpreting the X-ray image of the respective bag. This is a demanding task which requires certain aptitudes and abilities and specific training (e.g., Koller et al., 2008; Schwaninger et al., 2005). Deriving from findings regarding consideration information for digit recognition the question arose if providing consideration information can improve X-ray image interpretation competency by giving a decision aid. Bernstein and Li (2005) investigated the role of consideration information on a digit recognition task and found evidence that the consideration information did improve the participants' recognition performance. In this study it was assessed if screeners, when presented consideration information, achieve a higher detection performance for IEDs. The results indicate that this is not the case. There was no difference found in detection performance between the three experimental conditions Consideration Information, Feedback, and Control. This means, there is no advantage of the consideration information for the X-ray screening task, which is contrary to the finding of Bernstein and Li (2005). However, it has to be pointed out that the task is very different. Although it is about recognition, in the X-ray screening task there is much more information

to be taken into consideration (a passenger bag packed with many other objects), whereas in the digit recognition task only one digit was presented which had to be recognized. Even if the expert status (novice, intermediate or expert screener) is taken into account, there is no effect to be observed. No significant differences between the experimental conditions could be found for any status group, which is defined considering working experience or considering X-ray image interpretation performance level. The distinction between experts considering working experience and experts considering performance level seemed interesting because different criteria are provided for. Significant correlations were found for the relation between detection performance in this study and days on job and performance level, respectively, and that for the consideration information condition, which in this case interested most. These correlations say that the shorter the working experience and the higher the performance level, respectively, the better the detection performance when consideration information is provided. Since there could not be found an effect of experimental condition on detection performance for neither status group, other factors seem to play a role causing this effect. The most probable reason is that screeners with shorter working experience in general have higher X-ray image interpretation competency. This may very well be since in the last years job applicants have to pass an X-ray image test in the course of a pre-employment assessment. This signifies that all screeners with a working experience of less than four years feature a minimum level of X-ray image interpretation competency, which is not the case for screeners who were employed prior to the application of an X-ray image test in pre-employment assessment. Although there is no real baseline measurement which could confirm this assumption, the correlation for the control group points in the same direction. Another fact which has to be considered is that IED detection is mostly rather weak at the beginning and can be increased significantly with training (Schwaninger & Hofer, 2004). However, as Schwaninger and Hofer (2004) showed, after only two months of training detection performance of IEDs could already be increased considerably. So this argument plays no important role here. Another assumption, however, could be that younger people (hereby assuming that short working experience correlates with younger age, which, however, would have

to be investigated thoroughly in the future) are less reluctant towards the use of new features like consideration information. Regarding the positive correlation between detection performance in this study and X-ray image interpretation competency the relation is obvious. It can be assumed that the higher detection performance in the consideration information condition is due to the screeners' generally higher X-ray image interpretation competency. Again, the significant correlation between detection performance and X-ray image interpretation competency for the control group supports this assumption, although a real baseline measurement is lacking. The analysis regarding reaction time revealed that the consideration information condition group had significantly higher reaction times on average than the other experimental condition groups. This makes sense considering their use of the consideration information. However, seeing that detection performance with consideration information did not improve, it means that it is only a waste of time. Screeners are focusing on something else instead of the main image and losing time without having a benefit from it. This is consistent over the status groups when defined by working experience: novices, as well as experts and intermediate screeners show the longest reaction time in the consideration information condition. According to the ANOVA this is only significant for the expert and the novices group. However, the differences are very small compared to the other experimental conditions. When status is defined by X-ray image interpretation competency the pattern is the same, but the ANOVAs reveal a significant effect only for the intermediate screeners, which is exactly contrary to the results based on DOJ status. But since the effects are very small and only state what is logical, they are not interpreted further. Another analysis revealed that the number of clicks on the buttons to highlight the IED in the consideration information is considerably higher on average in the feedback condition than in the consideration information condition (in the control condition no consideration information was displayed at any time). But here as well, no benefit for detection performance resulted from it. When looking at the status groups, novices - when defined by working experience - in the feedback condition highlight the IEDs in the consideration information far more often than intermediate or expert screeners. However, this difference is not significant, which may be due to the large standard de-

viations. There appears to be a large difference between the screeners with regard to how often they use the buttons. Just interpreting this tendency it could be presumed that - again assuming that little working experience correlates with younger age - younger people are less reluctant towards new features and technological facilities and thus activate them more frequently. However, in the consideration information condition there is no difference between the three DOJ status groups regarding use of the buttons. Again, there are large differences between the screeners regarding the frequency of button activation. If analyzed by CAT status, the pattern is again similar to the analysis based on working experience. Here, the novices - defined by low X-ray image interpretation competency - are using the highlighting buttons more frequently in both experimental groups. The data show a steady increase in use of buttons from experts over intermediate screeners to novices. The differences between the status groups are smaller. In general, the tendency shows more frequent use with decreasing X-ray image interpretation competency. It could be assumed that screeners with higher competency are more "self confident" and therefore do not feel the need to use a decision aid. The selection of consideration information using PCA has been done with regard to the potential application. If there would have been a significant benefit of consideration information in the X-ray screening task it might have been worth considering its implementation. The idea would be that, whenever an X-ray image appears on the screen, the appropriate consideration information would be automatically selected by an algorithm and displayed on the screen. This algorithm would have to base on image features and pixel information. To sum up, according to the data in this study it seems that consideration information, presented simultaneously to the main image as well as in the form of feedback, only needs time to look at but has no benefit for detection performance whatsoever. At least, this holds for the IED detection. IEDs were chosen because there are large differences in detection performance between screeners and a large increase in detection performance can be achieved with appropriate, individually adaptive computer-based training (e.g., Koller et al., 2008). To definitively rule out any advantage of consideration information in the task of X-ray security screening further studies will have to be conducted. For example, in another study it would have to be

ensured more carefully that the screeners understand the instructions. The difficulty in such a study is that participants must not receive too much information regarding the aims in order not to distort the results. Furthermore, the supporting effect could only develop with recurrent use and increased familiarity of the tool. And last but not least, the so far only vague but not statistically significant effects could become manifest with a larger sample of aviation security screening officers.

Part III

Certification Analysis and Standard Setting

6

Different Ways of Analyzing Certification Tests for Aviation Security Screening Officers and Their Implications

There are many situations where people have to produce evidence of their knowledge, be it in school, at university or at work. This knowledge is usually assessed using some sort of test or exam. Widely used forms of tests are multiple choice tests and open-ended questions. However, there are as well areas in which more specific tests and practical tests, respectively, are likely. For example, in fields where the interpretation of any kind of image material is an essential component of the task, image-based tests can be applied. Thus, the optimal test format depends on the topic and the purpose of the test. In this paper we do not focus on the tests which should be applied, but discuss different kinds of analysis for the two test forms, multiple-choice tests and image interpretation tests in the field of aviation security. Aviation security screening officers have to be certified periodically in order to ensure and monitor the high level of security. The task of aviation security screening officers is to prevent individuals from bringing potentially hazardous objects on a plane which could endanger the safety and lives of other passengers and the crew. The usual procedure for reaching this goal is to subject all passengers and their baggage to a strict security check. While the passengers have to pass a metal detector and are hand searched when under suspicion of carrying a prohibited object, their baggage is being X-rayed. The screening officers see the X-ray images of the passenger bags on their monitor and have to search it for objects not allowed to be carried on the plane. Depending on their operational area screening officers have to solve image-based tests covering this area. Furthermore, a theoretical exam in the form of a multiple choice test has to be conducted. The first part of this paper, therefore,

focuses on multiple choice tests, their characteristics, strengths and weaknesses and discusses some issues regarding the scoring of multiple choice tests. The second part deals with image-based tests and their analysis.

6.1 Multiple Choice Tests

Multiple choice tests allow an efficient and economic assessment of theoretical knowledge by giving predefined response options. To be correct, multiple choice refers to questions where one of n response options has to be chosen as the correct answer (L. J. Gross, 1982). Oftentimes the term multiple choice is used for each kind of test with predefined response options to be marked. However, there are several alternatives of this form of test. For the sake of clarity the most common forms of multiple choice tests and their names as used in the remainder of this manuscript should be pointed out.

Table 6.1. Different forms of multiple choice tests.

Item format	Characteristic
Multiple Choice [MC]	$n > 2$ response options where only one is correct
True-false [TF]	$n = 2$ response options where one is correct
Multiple true-false [MTF]	$n > 2$ response options where more than one can be correct

Besides being more efficient and economic, the advantages of multiple choice tests, as compared to open-ended questions, are their higher reliability per testing hour and their less expensive costs of production. They also allow for standardized analysis. Following the general rule that the more items a test possesses the more reliable it is, multiple choice tests result in featuring a higher reliability per hour of testing time. This is because multiple choice tests, having only predefined response options which have to be marked, require less time to be solved, that is, more items can be solved

within an hour (Schuwirth & Vleuten, 2004). Open-ended questions, however, provide a very inefficient measurement. Not only do they require a large investment of examinee time to produce scores with acceptable reliability, they also yield complex responses which are time consuming and demanding to evaluate (Ward, 1982).

Besides their advantages, multiple choice tests bear some problems. In comparison with tests using a free response format, the question arises of whether the same or different measurement dimensions are being used. This is because experiments on learning have shown that examination of learnt materials is harder if reproduction rather than recognition of the learnt materials is required (Kubinger & Gottschall, 2007). Another problem is the guessing phenomenon which can occur when using a multiple choice response format for the test items. A test-taker who does not know the correct answer to a question nonetheless can select it by sheer luck with a non-negligible probability (i.e., 1/n, where n is the number of response options). If a test-taker has partial knowledge this probability can even be increased because some options can be eliminated with certainty. Many investigations of guessing on multiple choice tests have come to the generally agreed on conclusion that guessing contributes to error variance and diminishes the reliability and validity of tests (e.g., Angoff, 1989; Carroll, 1945; Frary, 1969; Glass & Wiley, 1964; M. E. Gross & Wright, 1985; Horst, 1933; Plumlee, 1952; A. O. H. Roberts, 1962; Zimmerman & Williams, 1982; Zimmerman, Williams, & Symons, 1984). There exists the dilemma between encouraging test-takers to answer all questions, independently of their knowledge, which introduces a source of random variance decreasing both reliability and validity, and discouraging test-takers from guessing, which causes the test results to be contaminated by personality factors regarding compliance (Rowley & Traub, 1977). The probability of guessing correct answers can in theory be reduced by increasing the number of response options (MacCann, 2004). However, this opinion is not shared by all authors. Rodriguez (2005), among others, recommends using only three options. According to some authors, examinees are unlikely to engage in blind guessing but rather eliminate what appear to them to be the least plausible distractors, thereby essentially reducing 4- or 5-option items to 3-option items (e.g., Costin, 1972, 1976; Kolstad, Briggs, & Kolstad, 1985). Kolstad et al. (1985) argue that the

quality of the distractors guards against guessing and not the number of distractors. However, as Rodriguez (2005) points out, this might only be valid for low speed tests but not for lower-ability students in high speed tests. According to Zimmerman and Williams (2003), who performed mathematical calculations on the effect of chance success, guessing is a nontrivial source of error in multiple choice tests, perhaps the largest in some tests, especially for tests with small numbers of items and/or response options. However, in many situations eliminating the guessing component of scores would increase reliability only at the greater expense of reducing validity (Frary, 1969). Besides the negative effect on reliability and validity, guessing effects indicate unfair assessment: the probability of passing the test is to a certain extent independent of the ability of the test-taker, but depends rather on the luck in guessing. This discriminates test-takers with moderate ability compared to test-takers with lower ability since a certain score can be achieved without any knowledge but only with lucky guessing (Kubinger & Gottschall, 2007). Furthermore, guessing has a reciprocal effect with item difficulty which in turn depends significantly on the conceptualization of different multiple choice response formats (Kubinger & Gottschall, 2007). Bar-Hillel, Budescu, and Attali (2005) present and criticize two ways in which test makers try to attempt to reduce the guessing problem, namely minimizing the incentives for guessing and reducing the opportunities for successful guessing. Usually, the approach to minimize the incentives is to penalize errors by subtracting points or fraction of points for each wrong answer. Minimizing the opportunities for guessing successfully can be achieved by avoiding clues in the response options (e.g., grammatical incompatibility of the option with the question, elaborating the correct answer in more detail than distractors, disproportional probability of a correct answer at a certain position, i.e. in the middle, etc.), which is referred to as key balancing. Bar-Hillel et al. (2005) criticize the key balancing method, instead of reducing cues for guessing, of actually introducing into answer keys a powerful cue to successful guessing, namely the negative dependencies between the positions of correct answers in successive questions. However, the use of computer-based tests allows for randomizing the appearance of the questions which would diminish the criticism of the key balancing method.

There exist several response-scoring modes which can be applied to multiple choice tests the aims of which are to improve on score characteristics over those obtainable by simply counting correct responses with no inhibition of guessing (number-right response-scoring method). Response-scoring is based on the assumption that examinees either know the answer to a test item or else choose among all alternative response options randomly (e.g., Diamond & Evans, 1973; Lord, 1975). According to Lord (1975) the advantage of formula scoring over number-right scoring depends on the number of omitted responses which in turn is depending on the knowledge level of the test-takers: the advantage will be negligible for high-ability students knowing the correct answers and be greatest for low-ability students omitting many items. Bar-Hillel et al. (2005) even recommend using number-right scoring. They criticize formula scoring as being more demanding, featuring incomplete, misleading, and unethical instructions, being self defeating, and introducing irrelevant variance. Formula scoring is adopted with the intention that it reduces the variability of the guessing-tendency variable in the test scores which in turn reduces test reliability (Cureton, 1966). However, mixed and even contradictory results with respect to reliability and validity tests resulting from the use of such response-scoring modes reported in the literature are found fault with (Frary, 1982). Typically, in MC questions (see Table 6.1) one point is awarded for a correct response and zero points are awarded for an incorrect response. One prominent procedure to deal with the guessing problem in multiple choice tests is the correction for guessing (CFG) which is used to weight response options. It is calculated with the formula

$$S = R - W/(k-1)$$

where S is a corrected score, R is the number of items marked correctly, W is the number of items marked incorrectly, and k is the number of response options for each item. Using this formula the examinee's score is corrected for the effects of guessing. This equation underlies the assumptions that all wrong answers are guessed wrong and that all correct answers are obtained either by knowledge or by guessing. No consideration is given to misinformation and partial information (Diamond & Evans, 1973), some authors criticize CFG because it overcorrects for misinformation and

undercorrects for partial information (e.g., Hammerton, 1965; Little & Creaser, 1966; Rowley & Traub, 1977).

Muijtjens, Mameren, Hoogenboom, Evers, and Vleuten (1999) obtained results in their study indicating that partial knowledge is measured more completely when number-right scoring is used. Davis (1964) and Gronlund (1965) seem to agree that the amount of overcorrection and undercorrection, respectively, depends on the particular test and the examinees. With CFG, correct answers receive a weight of 1, incorrect answers receive a weight of -1/(k-1), and omissions receive a weight of 0. Several studies (Sherriffs & Boomer, 1954; Votaw, 1936) have shown that the CFG results in certain types of examinees being penalized more than others on the basis of irrelevant personality factors (Frary, 1969; L. J. Gross, 1982). However, there is no clear agreement on what influences the extent of guessing more: knowledge or personality factors (Diamond & Evans, 1973). Cross and Frary (1977) recommend assuring that clear test directions be formulated and that test-takers be taught effective test-taking behavior before adopting conventional correction-for-guessing. For TF (see Table 6.1) items the effects of guessing are apt to be great since the chance of guessing correctly is 50%. Under the TF test format the CFG formula reduces to (R-W). The MTF format as a special case of both the MC and the TF format gives credit to partial knowledge since the examinee is allowed to make a judgment on every option independently of the others. With MTF the question on scoring arises: should each item be scored as an extension of the MC type, whereby the item is scored as one entity, or should it be scored as an extension of the TF type, in which each option is scored independently. The latter provokes the issue of whether CFG should be utilized (L. J. Gross, 1982). Most research articles discussing MTF scoring techniques advocate the application of CFG with the aim of minimizing or even eliminating the effects of guessing (e.g, Buckley-Sharp & Harris, 1971; Dugdale, 1971; Harden, Lever, & Wilson, 1969; Lennox, 1967; Sanderson, 1973). Generally, with MTF those options are required to be marked that are believed to be true and no mark is given to options believed to be false. One point is awarded for each correct mark and one point is subtracted for each incorrect mark. Thereby there is no possibility to distinguish between a false opinion and an omission. L. J. Gross (1982)

recommends not applying CFG on MTF tests because the score range is in favor of being sufficiently large for adequate variability and discrimination without it and factors unrelated to ability or achievement that lower validity may be introduced. L. J. Gross (1982) postulates that examinees should be instructed to mark each response option to be true or false and the score being only the number of correct responses. This strategy eliminates the effects of subjectivity in response style and allows an adequate discrimination between students.

Much research has been conducted on the advantages, disadvantages, and differences of various multiple choice test formats (cf. Table 6.1). Frisbie (1992) conducted a review especially on MTF items and comes to the conclusion that the measurement of educational achievement could be improved if MTF items were used more widely be it for program placement, classroom assessment, or certification of competence. MTF items automatically assess partial knowledge via the answers provided to each true-false option. It is only a matter of scoring method whether partial information is taken advantage of or not (Albanese & Sabers, 1988). Findings relating formula scoring and test reliability are very mixed and even contradictory. In part, this inconsistency is undoubtedly based upon the extreme variation in the character of the tests and populations involved (Frary, 1982). Test instructions regarding guessing have a crucial effect on the reliability depending on the scoring mode as well. Studies investigating instructional patterns (e.g., no comment on guessing, encouragement to guess with no scoring penalty, "do not guess" instructions with scoring penalty) and their relative scoring (e.g., rights-only, corrected scores) revealed different results regarding reliability. Most of the studies found higher reliabilities for uncorrected scores (e.g., Glass & Wiley, 1964), however, some found no difference in reliability estimates subject to scoring methods (Diamond & Evans, 1973) or even higher reliability for formula scoring (e.g., Muijtjens et al., 1999). According to Mattson (1965) and Magnusson (1967) random guessing reduces reliability by increasing error variance without a corresponding increase in true score variation. Brandenburg and Whitney (1972) found in their study, comparing different scoring methods for matched pair true-false item tests, larger internal consistency coefficients for scoring methods having larger effective score ranges. Waters (1976), under the assumption

that partial knowledge is defined as the selection of an option through the elimination of one or more alternatives as incorrect or implausible, hypothesizes that a scoring system which assesses the partial knowledge of an examinee would increase the reliability and validity of the test.

Tests commonly used in educational and psychological measurement feature the property that the longer they are the more reliable are the scores they yield (Ebel, 1969, 1972). Zimmerman and Williams (2003) emphasize an interaction between the number of items of a test and number of response options for the items as determinants of the reliability (see also MacCann, 2004). According to these authors, increasing the number of response options strongly influences the reliability of relatively short tests, whereas for relatively long tests increasing the number of response options has a smaller effect on the reliability. Their practical message is to use five-choice items for short tests whenever possible, or in other words, if a small number of choices per item is used, it is essential to make the test as long as possible. Haladyna and Downing (1989) recommend to "develop as many functional distractors as are feasible", that is, the key is not the number of options but the quality of distractors. However, Haladyna, Downing, and Rodriguez (2002) and Rodriguez (2005) are even of the opinion that three response options are sufficient in most instances arguing that the effort of developing the fourth option is probably not worth it. They state that using more options does little to improve item and test score statistics and often results in implausible distractors. Rodriguez (2005) points out though, that the impact of changing the number of options depends on the method of deleting options; if they are deleted randomly, say from 4- to 3-options, this can have a detrimental effect on reliability. However, test reliability is not affected if only ineffective distractors are deleted.

Regarding the validity of a test, depending on test instructions and scoring, the tendency was found that validity is slightly higher for corrected scores and for the "do-not-guess" instruction. However, as with reliability, results are not clearly without ambiguity and the opinions of authors regarding the use of correction and the instructions are not unanimous (Diamond & Evans, 1973). Contradictory results

have been reported on whether the use of CFG improves test validity (e.g., Cureton, 1966) or if it fails to increase test validity and in comparison with number-right scoring features even lower validity despite the fact of higher KR-20 reliability (Jaradat & Sawaged, 1986). Validity should be enhanced whenever the criterion measure is sensitive to partial knowledge (Frary, 1980).

6.2 Image-based Tests

Image-based tests are applied to measure the ability of aviation security screening officers to detect prohibited items in passenger bags and to discriminate them from harmless objects. Economic reasons, among others, require the screening officers to do a very efficient job. That is, the time used per passenger for the security check should be as short as possible, but without compromising the security. This should also minimize the waiting queue at the security checkpoints. By screening the bags with X-rays it is not necessary to open each bag and manually search it. This should be reduced to bags that seem to contain a prohibited item or bags that cannot be interpreted from the X-ray image due to their complexity or low transparency. Therefore, the screening officers have to exhibit a high ability to detect prohibited items and to discriminate them from harmless everyday objects. This ability can also be termed sensitivity in the sense of signal detection theory. With the X-ray image-based tests this sensitivity should be measured for certification of aviation security screening officers. Image-based tests are typical members of yes-no experiments, belonging to the one-interval design. The screening officers are presented with a single stimulus on each trial (i.e., an X-ray image of a passenger bag), drawn from one of two stimulus classes. In this case, the two stimulus classes are the passenger bags containing only harmless objects (Noise trial) and the passenger bags containing additionally a prohibited item (Signal-plus-Noise trial). The yes-no experiment measures discrimination, that is, the ability to tell the two stimulus classes apart. If, in this case, one stimulus class contains only the null stimulus (Noise trial), the task is called detection (Macmillan & Creelman, 1991). Therefore, we are also talking of detection performance of screening officers, meaning the sensitivity or

the ability to detect prohibited items and tell them apart from harmless items. As the test concerns two kinds of stimuli, Noise and Signal-plus-Noise trials, and two possible responses, OK (the bag does not contain any prohibited item) and NOT OK (the bag contains a prohibited item), any of four joint events can occur on a trial. Correctly detecting a prohibited item is called a hit; failing to detect it is called a miss. Mistakenly detecting a harmless object as prohibited is a false alarm; correctly responding OK to a Noise trial is called a correct rejection (Macmillan & Creelman, 1991). The proportion of all Signal-plus-Noise trials to which a person responded NOT OK is the hit rate (H); the proportion of all Noise trials to which a person responded NOT OK is the false alarm rate (F). It would not make much sense to just count all hits in order to assess the detection performance of a screening officer. A high hit rate could easily be achieved by simply responding NOT OK to all trials. This would generate a hit rate of 1 (all bags containing a prohibited item have been judged correctly) completely ignoring the high false alarm rate of 1 as well (all bags without any prohibited item have been incorrectly judged as containing one). At the airport security checkpoint, this would entail opening all bags and thus unnecessarily generate a long waiting queue. Therefore, the false alarm rate is an indicator of a person's efficiency. The sensitivity or detection performance of a person taking all bags out would be zero and the detection of prohibited items would be by chance instead of sensitivity. So for a reasonable measure of detection performance the false alarms also have to be taken into account. Sensitivity equals a specific measure of the discrepancy between a hit rate and a false alarm rate (Macmillan & Creelman, 1991). A common way to achieve this is by applying signal detection theory, generating measures as for example *d'*. The sensitivity measure *d'* is defined in terms of z, the inverse of the normal distribution function: $d' = z(H)\text{-}z(F)$ (Green & Swets, 1966). Until now, the image-based tests used for certification of aviation security screening officers have been analyzed with signal detection measures. However, this approach is not immune to guessing and incorporates a certain inaccuracy in the measurement. It is based on the assumption that when responding with NOT OK the person has detected the prohibited item correctly. However, this could be a fluke because the person could have judged a harmless object as prohibited and actually missed the

real prohibited item. Nevertheless, it would have been a hit. To measure whether there is a large difference between judging a bag as NOT OK from the actual threat item or from a harmless object, screener officers have to identify the prohibited item by marking it on the screen. If analyzed according to this data, the true hits can be assessed, namely, only those hits where the target (i.e., the prohibited item) has been identified. This definition raises the question of how the other response possibilities are defined (the Noise trials still generate false alarms and correct rejections). With the instruction to mark the threat object in the image two kinds of flukes for hits can result: either a wrong (i.e., harmless) object was marked or nothing was marked at all. The first case we call false click, the second one false hit. Intuitively, the false hits and false clicks would count as misses, as do the Signal-plus-Noise trials that have been judged as OK. The false clicks could just as well be counted as false alarms though (the person marked a harmless object as prohibited). For the pure assessment of a screening officer's detection ability of prohibited items this question is of no consequence since the real hits serve as a good measure. However, there are two critical points. First, this approach does not eliminate the guessing problem completely; a screening officer can still mark a dubious object by luck (probably because it looks suspicious even if this person cannot define the object). Second, although the detection ability can be measured quite well using only the true hit rate, there is no information in this data about the efficiency of a screening officer. Still the false alarm rate has to be taken into account in order to assess the sensitivity, that is, the ability to discriminate harmless from prohibited objects. This however raises the question on which values are to be used for signal detection measures. What is the definition of a false alarm? Unfortunately no literature can be found on this specific topic within signal detection theory. This study examines how big the difference in the hit rate, on the one hand, and in the detection performance values, on the other, will be when calculating with all hits and with true hits, respectively. If a difference can be found the question of the meaning of these results arises and ultimately the one which measure should be used to assess the sensitivity of the aviation security screening officers. To simplify matters, false hits and false clicks are counted as misses for the true hit analysis within the scope of this study.

6.3 Experiment 1

The first experiment of this study deals with the scoring of a theoretical Multiple True-False (MTF) test using a theoretical exam for aviation security screening officers as an example. The question is how different scoring modes affect difficulty and reliability of the test and which one should be used. A further focus of investigation lies on the instruction and its effect on the difficulty level and reliability of the test. One data set is based on no information for the test-takers on how many response options are correct, the second data set is based on the information if an item has 1 out of n correct options or n out of n.

6.3.1 Method

Participants

In the course of their yearly certification 192 aviation security screening officers from a European airport conducted the theoretical exam on computer (TEC). The screening officers conducted the TEC under one of two conditions. One group of aviation security screening officers (n = 78) received the information whether only 1 or whether n of n response options were correct. The other group (n = 114) received no information about the number of correct response options.

Materials

The computer-based theoretical exam (TEC) consists of two main parts. One part includes a number of questions which are defined by the national appropriate authority. The second part is adapted to each individual airport with airport specific questions. These questions are defined by the airports themselves. The version of the TEC used in this study contained 40 questions of the appropriate authority. The questions cover all job aspects, from knowledge about the actual task of security screening over administrative issues up to regulations. The questionnaire consists of MTF items where 1 up to n of n response options can be correct. Table 6.2 gives an

overview on the number of items depending on their number of response options, including the information on how many of them are correct, only one or more than one. The questions of the appropriate authority served as database for the following analyses. The questions appeared in random order and the examinees had to answer the questions by marking the correct response options and not marking the wrong response options, respectively. Marks could be deleted by clicking on the box again until the question was submitted. A once submitted question could not be repeated.

Table 6.2. Number of questions per number of response options in the questionnaire. Also indicated is the number of the respective questions with one correct response option and the number of questions with more than one correct response options, respectively.

Number of response options	Number of questions in questionnaire	Number of questions with n = 1	Number of questions with n > 1
3	14	13	1
4	17	11	6
5	7	2	5
6	2	1	1
TOTAL	**40**	**27**	**13**

Procedure

The aviation security screening officers conducted the theoretical exam in the course of their periodic certification. One group of the examinees was informed if only one response option was correct or if n response options were correct, however, without any information on the exact number. The second group received no information about how many response options were correct for the different questions. The test results were analyzed with four different scoring methods. The following scoring methods were applied:

Multiple Response (MR) (Albanese, Kent, & Whitney, 1979): one item is treated as an entity and one point is awarded only if all the response options of an item were marked correctly. In the case of this study not marking a false response option counted as correct answer (neglecting the problem of omissions due to lacking

knowledge). Harasym, Norris, and Lorscheider (1980) referred to this scoring method as dichotomous scoring. In the remainder of this paper it is always referred to as MR. The MR scoring method is based on the assumption that reduced probability of chance success results in increased reliability of the test (Frary & Zimmerman, 1970). The reduced probability of chance success is suggested to be achieved through longer tests and more response options per multiple-choice item (Ebel, 1969; Wesman, 1971). A potential weakness of the MR method is its ignorance of partial knowledge. Any useful information contained in the score of a person answering only some but not all response options correctly is ignored. Answering some but not all of the response options correctly counts as if none were answered correctly. There is evidence that factoring partial knowledge into test scores will lead to increased test reliability and validity. However, it is questionable as to whether the effort of applying such complicated procedures is worth the gain (Albanese & Sabers, 1988).

Scoring method #2: Unlike the MR method, which is an all-or-nothing proposition, the second scoring method takes into account some partial knowledge by apportioning credit for beyond chance success levels but awarding no credit for performance at or below chance levels. In the case of an item with 4 response options $\frac{1}{2}$ credit would be awarded for three options correct and full credit for all options correct (Albanese & Sabers, 1988).

Scoring method #3: The third scoring method rewards even more partial knowledge, namely performance at chance level. In the case of an item with 4 response options $\frac{1}{3}$ credit would be awarded for 2 options correct (chance level), $\frac{2}{3}$ credit would be awarded for 3 options correct, and full credit would be awarded for all options correct (Albanese & Sabers, 1988).

Multiple true-false (MTF) (Albanese et al., 1979): The MTF scoring method apportions credit evenly for all true-false options answered correctly. The basic principle underlying MTF is the assumption that each response an examinee makes contains valid information, even for performance below chance levels. It is important to distinguish the MTF item format (see Table 6.1 above) and the MTF scoring method. Table 6.3 shows the distribution of credits awarded for each correctly answered response option depending on the scoring method and the number of response

options per item. As mentioned before, a correctly marked response option, as well as a correctly unmarked response option, counts as correct answer.

Table 6.3. Credits awarded for a certain number of correct response options depending on scoring method and number of response options per item. MR = multiple response, Method #2 = scoring method 2, Method #3 = scoring method 3, MTF = multiple true-false.

		Scoring method			
		MR	Method #2	Method #3	MTF
Item with 3 response options	1 correct	0	0	0	$\frac{1}{3}$
	2 correct	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$
	3 correct	1	1	1	1
Item with 4 response options	1 correct	0	0	0	$\frac{1}{4}$
	2 correct	0	0	$\frac{1}{3}$	$\frac{2}{4}$
	3 correct	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
	4 correct	1	1	1	1
Item with 5 response options	1 correct	0	0	0	$\frac{1}{5}$
	2 correct	0	0	0	$\frac{2}{5}$
	3 correct	0	$\frac{2}{4}$	$\frac{1}{3}$	$\frac{3}{5}$
	4 correct	0	$\frac{3}{4}$	$\frac{2}{3}$	$\frac{4}{5}$
	5 correct	1	1	1	1
Item with 6 response options	1 correct	0	0	0	$\frac{1}{6}$
	2 correct	0	0	0	$\frac{2}{6}$
	3 correct	0	0	$\frac{1}{4}$	$\frac{3}{6}$
	4 correct	0	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{4}{6}$
	5 correct	0	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{5}{6}$
	6 correct	1	1	1	1

Test difficulty, that is, mean scores, and reliability values are calculated and compared for each scoring method and between the two groups.

6.3.2 Results

The test scores for each screening officer are calculated according to Table 6.3. Individual data were averaged across all examinees within one group (with vs. no information) for each scoring method in order to obtain the mean test score and test difficulty, respectively. Figure 6.1 compares the mean test scores obtained by the calculation according to the four different scoring methods and between the two groups. Independently of the information level, that is, whether the candidates had the information that one or more than one response options are correct or not, the mean test score is lowest for the multiple response (MR) scoring method and highest for the multiple true-false (MTF) scoring method. Mean score for scoring method 3 is slightly higher than for scoring method 2 but the increase is smallest between these two scoring methods. Independently of the scoring method used, the mean test scores are higher for the group which was informed that one or if more than one response option is correct compared to the group which received no information regarding the number of correct response options per question. However, the difference in mean scores between the two groups is largest for the MR method and smallest for the MTF method.

A two-factor analysis of variance (ANOVA) on mean percentage test scores with group (with vs. no information) as between-participants factor and scoring method (MR, method 2, method 3, MTF) as within-participants factor, revealed a significant main effect for scoring method with a large effect size of $\eta^2 = .94, F(3, 570) = 2990.61, p < .001$. The main effect of group was also significant and large with $\eta^2 = .20, F(1, 190) = 47.26, p < .001$ as well as the interaction between method and group: $\eta^2 = .21, F(3, 570) = 51.65, p < .001$. All effect sizes are interpreted according to the conventions of Cohen (1988). Figures 6.2 and 6.3 show the Cronbach's alpha and Guttman split half reliability values, respectively, comparing the four scoring methods and the two groups.

The Figures 6.2 and 6.3 reveal a substantially higher reliability for the test scores achieved when candidates had no information regarding the correct number of response options, regardless of the scoring method. Comparing the four different scor-

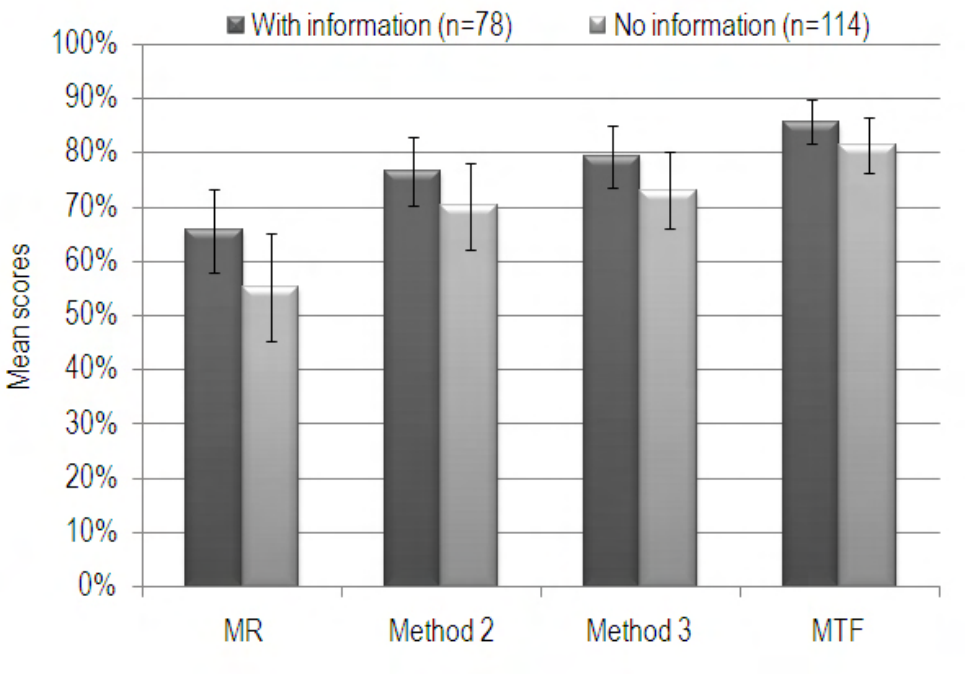


Fig. 6.1. Mean percentage scores for the two groups (with information vs. no information) for the four scoring methods: multiple response (MR), scoring method 2 (Method 2), scoring method 3 (Method 3), and multiple true-false (MTF). The thin bars represent the standard deviation.

ing methods, the reliability values amount on a comparable level, independently of the mean test score a scoring method yields (cf. Figure 6.1).

6.3.3 Discussion

Looking at the percentage correct scores for the different scoring methods, the findings of Albanese and Sabers (1988) could be replicated confirming that incorporating more partial information increases test scores. These authors also found the lowest score for MR method and the highest score for MTF scoring with scoring method 2 and 3 lying in between. Despite the advantages of MTF over the other scoring methods studied, several concerns are expressed concerning the use of MTF scoring. First, compared with the other scoring methods, MTF tends to produce relatively high percentage correct scores. Second, partial knowledge is not always worth being rewarded (think of a physician not treating a disease correctly). Third, in some circumstances reduction of chance success may be the overriding concern, for example

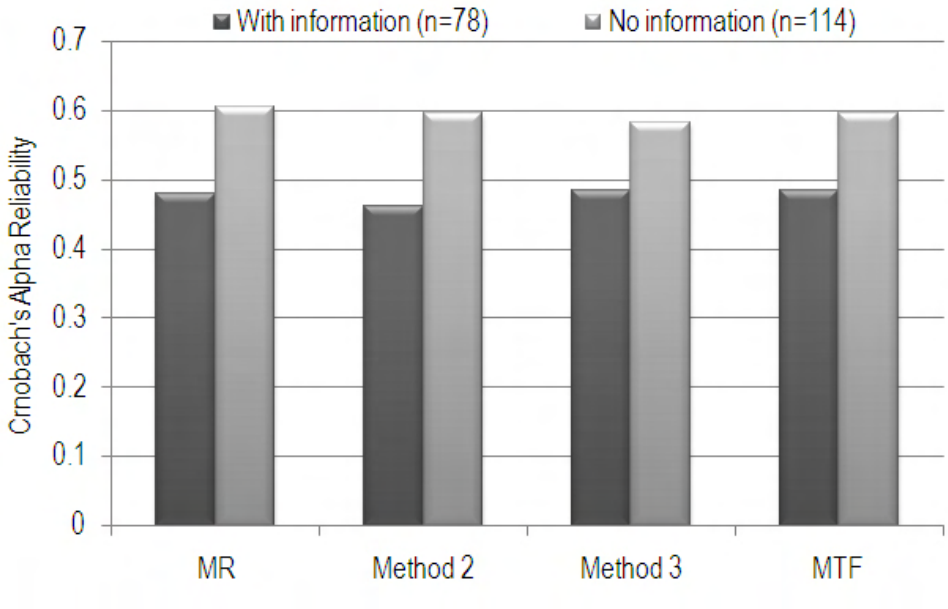


Fig. 6.2. Cronbach's alpha reliability values for the four scoring methods: multiple response (MR), scoring method 2 (Method 2), scoring method 3 (Method 3), and multiple true-false (MTF) and the two groups (with information regarding the number of correct options vs. no information regarding the number of correct options) separately.

for certification examinations, where the intention is to only certify personnel with the appropriate knowledge level. The last two concerns could be best overcome using MR scoring or one of the intermediate formulas, the scoring method 2 and 3 (Albanese & Sabers, 1988). These concerns imply the need to define the concerns and demands for each domain individually. In the case of the theoretical certification of aviation security screening officers awarding partial knowledge can be regarded as reasonable, which speaks for the MTF scoring. However, a combination of the two methods depending on the question should also be taken into consideration. That is, questions which do not allow for partial knowledge and others in which partial knowledge should be awarded. Contrary to the results of Albanese and Sabers (1988) and Frisbie (1973, 1992), we found no significant change in reliability for the different scoring modes in our study. Albanese and Sabers' (1988) data leads to the conclusion that the last three scoring methods shown in Table 6.3 are essentially equivalent to each other and superior to MR scoring in terms of reliability. This is a contrary finding to the assumption that reduced probability of chance success results

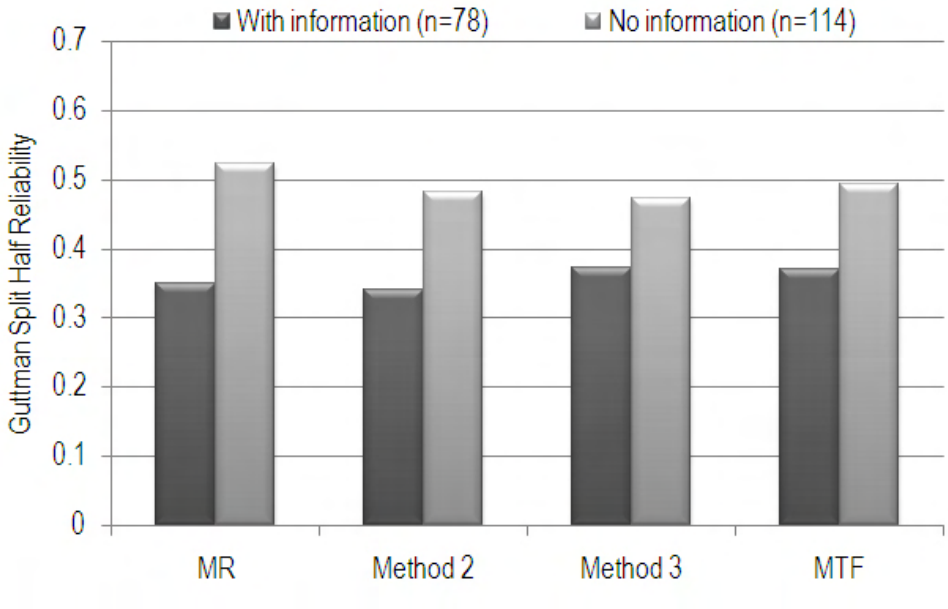


Fig. 6.3. Guttman split half reliability values for the four scoring methods: multiple response (MR), scoring method 2 (Method 2), scoring method 3 (Method 3), and multiple true-false (MTF) and the two groups (with information regarding the number of correct options vs. no information regarding the number of correct options) separately.

in increased reliability of the test (Frary & Zimmerman, 1970). Frisbie (1973, 1992), in his comparison of MC versus true-false items and his review on research about MTF, confirmed the tendency of higher reliability for MTF items than for MC (cf. Table 6.1) or true-false item and other objective formats. However, no significant change in reliability can be denoted for the different scoring modes in our study. This would deny Waters' hypothesis which says that, under the assumption that partial knowledge is defined as the selection of an option through the elimination of one or more alternatives as incorrect or implausible, a scoring system which assesses the partial knowledge of an examinee would increase the reliability of the test (Waters, 1976). Waters investigated tests with a slightly different response format than the one used in this study. There, the students did not have to mark all correct response options, but rather had a selection of different combinations of correct response options from which they had to pick the correct one[1].The author found higher internal

[1] Example: For a question with the answers A, B, C, and D, the response options could be as follows: 1) B only, 2) D only, 3) A and C, 4) A, C, and D, 5) A, B. and D.

consistency reliability if response options were scored that included the correct responses (e.g., if a student picked the response "A and C" while the options A, C, and D are correct) as compared to rights-only scoring. For the two tests Waters used which the response format is reported for, reliability, based on the point-biserial weighting, increased to .783 and .833, respectively, from reliability values based on rights-only scoring of .520 and .615, respectively. The difficulty levels of the tests used in Waters' study and those of the tests in this study are at a comparable as well (Waters, 1976: $p = 0.563$ and $p = 0.653$). Regarding information level, the group which was informed that a question had one or more correct options yielded higher percentage correct scores. However, this has a detrimental effect on the reliability, which is considerably lower for the data of this group. The reliability data therefore argue towards giving test takers no information regarding the number of correct response options which can be explained as follows. High performers can solve an item correctly regardless whether they have the information about the number of correct responses or not. They are capable to evaluate each response correctly whereas a low performer rely his / her response on the given information. Thus, the no information method seems to discriminate better between the individual test takers which results in higher reliability scores as well. Further, the difference in percentage correct scores indeed is significant; however, reliability values have a higher impact in the discussion about instructions. While the reliability is an innate feature of a test, a higher or lower mean percentage correct score can be compensated or taken account of by a lower or higher pass mark.

6.4 Experiment 2

The second experiment is on scoring image-based tests. To assess the competency of aviation screening officers different computer-based X-ray image interpretation tests are reasonable. In the applied X-ray screening tests, screening officers see X-ray images of passenger bags on the screen and have to judge if these bags contain a prohibited item or not. The images are of the same kind as those appearing on the monitor at the security checkpoint, where passengers have to pass the security

control and their baggage is being scanned with X-ray. During the test, the screening officers have to give the answer OK if the respective bag does not contain any prohibited item and the answer NOT OK if the respective bag contains a prohibited item. Furthermore, if a prohibited item was detected, the screening officer has to mark it by clicking on it on the monitor. This study examines whether it makes a difference in the overall performance if the test items were scored solely according to the judgment OK or NOT OK - and therefore to wrong or right and using signal detection measures - or if test items containing a prohibited item are only counted as a correct answer if the prohibited item was marked correctly. The first approach ignores the problem of the false hits and false clicks. If a bag containing a prohibited item was judged as NOT OK it counts as a correct answer independently of whether the actual prohibited item was detected or a harmless object was wrongly interpreted as prohibited and the actual prohibited item in fact was missed. However, the second approach only counts the true hit scores. As in Experiment 1, the aim is to find differences in difficulty and reliability of the tests depending on the scoring method and to come to a conclusion which one of the two methods is appropriate to use. Furthermore, the question arises on how to calculate signal detection measures with the true hit score.

6.4.1 Method

Participants

In total, 912 screening officers conducted the certification exam which constitutes the data basis for this experiment. The certification consists of the following four tests: the X-Ray Object Recognition Test (X-Ray ORT), the X-Ray Competency Assessment Test (X-Ray CAT), the X-Ray Bomb Detection Test (X-Ray BDT) and the Theoretical Exam (TEC). The X-Ray ORT and the TEC have to be solved by every screening officer in the course of their certification independently of their working domain. The X-Ray CAT has to be solved by screening officers who work in cabin baggage screening (n = 817) and the X-Ray BDT by screening officers who work in hold baggage screening (n = 155).

Material

The screening officers had to solve two or three X-ray image interpretation competency tests, depending on their working field. All tests are computer-based. The X-Ray ORT is used to measure the competency of interpreting X-ray images and dealing with factors like rotation and superposition of an object and complexity of an image in object recognition. It is often used as a pre-employment assessment test for aviation security screening officers because it measures mainly stable aptitudes and therefore is a good tool for the selection of qualified personnel. It consists of 256 X-ray images of passenger bags. Each bag is used twice; therefore 128 X-ray images of bags are used. Once each bag is presented as Noise trial that is, without any additional prohibited item. Once each bag is presented as Signal-plus-Noise trial, that is, containing one prohibited item. The prohibited items are knives and guns. Since the X-Ray ORT is designed to measure the abilities of a person to interpret X-ray images without requiring the domain specific knowledge, only guns and knives are used. It is assumed that every person knows what a gun or a knife looks like. Furthermore, the images are depicted in black and white since coloring of material in X-ray incorporates information that is not implicitly known. This test has to be solved by every screening officer in the course of the yearly certification. So each examinee sees each of the 128 suitcases twice, once as harmless bag and once containing either a gun or a knife. Figure 6.4 shows an example of the stimuli. The items are depicted on the monitor in random order and disappear after four seconds. Within these four seconds the prohibited item, if one was there and has been detected,

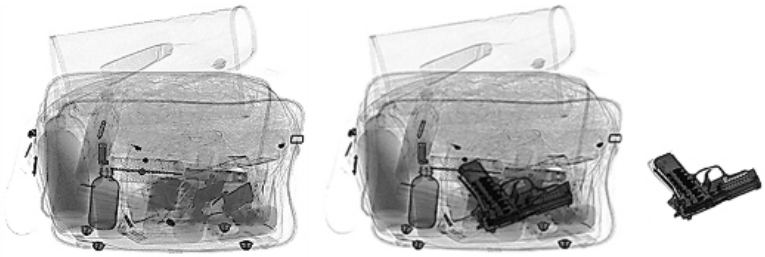


Fig. 6.4. Noise (left bag) and Signal-plus-Noise (right bag) trial of the X-Ray ORT. The prohibited item (gun) contained in the Signal-plus-Noise trial is depicted on the right.

has to be marked. For the other responses (OK or NOT OK and the difficulty rating) the time is unlimited. For more details on the X-Ray ORT and further research see for example Hardmeier, Hofer, and Schwaninger (2005) and Hardmeier et al. (2006a).

Fig. 6.5. Noise (left bag) and Signal-plus-Noise (right bag) trial of the X-Ray CAT. The prohibited item (taser) which is contained in the Signal-plus-Noise trial in the top left corner of the bag is depicted on the right (X-ray image on the top, real photograph on the bottom).

The X-Ray CAT not only assesses the ability to cope with X-ray images but also the knowledge of which items are prohibited and what they look like in X-ray images. Therefore, not only guns and knives but also improvised explosive devices (IEDs) and other prohibited items (like grenades, electric shock devices, self-defense gas sprays, etc.) are included. Figure 6.5 shows and example of the stimuli. The structure is the same as for the X-Ray ORT (256 items whereof 128 are Noise and 128 are Signal-plus-Noise trials) except that the items are presented in color for 15 seconds and the prohibited items belong to one of the four categories guns, knives, IEDs, and other threat objects. The X-Ray CAT has only to be solved by screening officers working in cabin baggage screening. For more information on the X-Ray CAT see Koller and

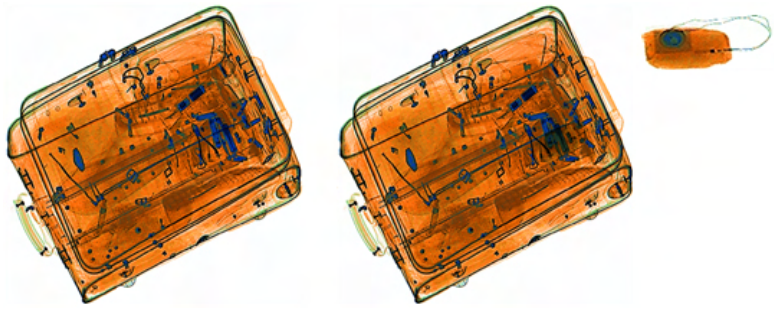


Fig. 6.6. Noise (left bag) and Signal-plus-Noise (right bag) trial of the X-Ray BDT. The improvised explosive device (IED) contained in the Signal-plus-Noise trial on the right side is depicted on the right of the bag.

Schwaninger (2006). The task is the same for all three tests: the examinee has to search each bag for a prohibited item and judge if one is existent or not. If an examinee detects a prohibited item, it has to be marked by clicking on it with the mouse and the bag has to be judged as NOT OK by marking the corresponding button. If no prohibited item was detected, the response OK has to be given. Furthermore, the difficulty of the item has to be rated on a slider bar[2]. In the X-Ray CAT, in addition to the identification of the prohibited item, the OK or NOT OK response, and the difficulty rating, the category to which the prohibited item belongs to has to be judged[3]. Except for the identification of the prohibited item, which has to be done while the image is on the screen, all answers can also be given after the image has disappeared. Figure 6.7 illustrates the user interface for the image-based tests.

Fig. 6.7. User interface of the X-ray image tests. This example illustrates a fictional screenshot from the X-Ray CAT.

6.4.2 Results

Before being subjected to analyses, the data was cleaned from data files where only 2 or less true hits were achieved. Most of these data files contained no true hits at all. The assumption is that corresponding candidates did not perform the task of marking the prohibited item. For the X-Ray ORT, data elimination was effected for

[2] These data are not included in the analyses in this study.
[3] These data are not analyzed within this study either.

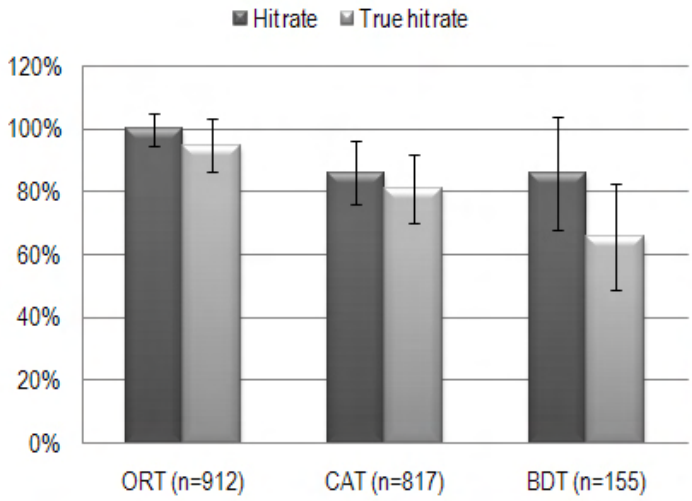
0.032% (n=30), for the X-Ray CAT for 0.089% (n=80) and for the X-Ray BDT for 0.025% (n=4) of all data files. This results in a final sample of 912 aviation security officers for the X-Ray ORT, 817 for the X-Ray CAT, and 155 for the X-Ray BDT.

Detection Performance

Measures of detection performance hit rate and *d'* were calculated. Both these measures were once calculated with uncorrected values and then compared to the ones calculated with the corrected values. The uncorrected values are the hit rate based on the response OK or NOT OK, regardless of whether the actual threat item was detected or a harmless object was mistakenly interpreted as a prohibited one, and the *d'* value calculated with this hit rate. Corrected values means taking into account only the true hits, that is, those hits where the candidates actually marked the target (the prohibited item). The false clicks, that is, the hits that resulted from correctly responding NOT OK to a test item but marking the wrong object (i.e., a harmless one) and the false hits, that is, the hits where nothing was marked at all, are counted as misses. This means, with this approach the false alarm rate remains the same for both procedures and only the hit rate is affected by this definition. Figure 6.8 shows the hit rate and the true hit rate and their standard deviations for all tests. Individual paired-samples *t*-Tests revealed significantly lower mean scores for the true hit rate for all three tests (all $p < .001$, see Figure 6.8 for values). The correlations between hit rate and true hit rate are significant as well for all three tests (all $p < .001$, see Figure 6.8 for values). Note: for security reasons values of all analyses have been multiplied with an arbitrary constant.

Figure 6.9 shows the detection performance *d'*, once calculated with the uncorrected hit rate and once calculated with the true hit rate. Individual paired-samples *t*-Tests revealed significant differences in *d'* for all tests (see Figure 6.9 for values). The correlations between the uncorrected and the corrected *d'* scores are significant for all tests (see Figure 6.9 for values).

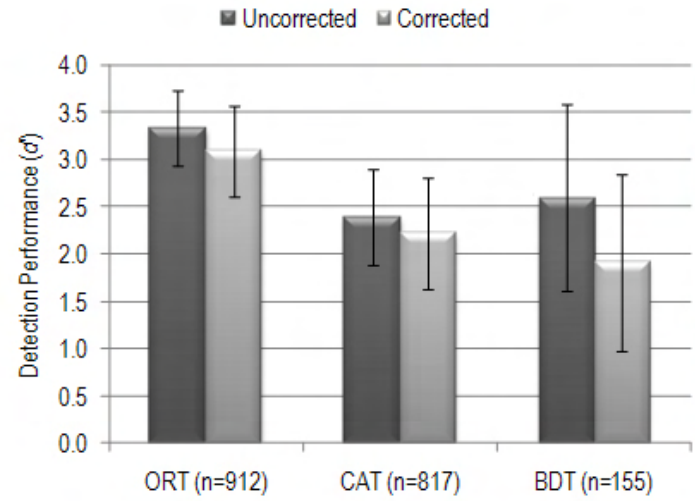
An independent samples *t*-Test with the false alarm rate of the X-Ray CAT and the X-Ray BDT showed no significant difference between the two false alarm rates



Test	Difference %	*t*	*p*
ORT	4.46%	21.93	< .001
CAT	4.56%	27.25	< .001
BDT	17.82%	21.71	< .001

Test	*r* Hit rate - True hit rate	*p*
ORT	0.668	< .001
CAT	0.902	< .001
BDT	0.829	< .001

Fig. 6.8. Hit rate (dark bars) and true hit rate (light bars) for the X-Ray ORT, X-Ray CAT, and X-Ray BDT with standard deviations (thin black bars). The results of the paired-samples *t*-Tests are depicted in the table on the top right. The correlations between the hit rate and the true hit rate for each test and the significance levels are depicted in the table on the bottom right.



Test	Difference *d'*	*t*	*p*
ORT	0.214	29.53	< .001
CAT	0.151	28.72	< .001
BDT	0.609	19.17	< .001

Test	*r d'* uncorr - *d'* corr	*p*
ORT	0.894	< .001
CAT	0.976	< .001
BDT	0.916	< .001

Fig. 6.9. Detection performance *d'* for the X-Ray ORT, X-Ray CAT, and X-Ray BDT, once calculated with the uncorrected hit rate (dark bars) and once calculated with the true hit rate (light bars). Thin bars represent standard deviations. The results of the paired-samples *t*-Tests are depicted in the table on the top right. The correlations between the uncorrected and the corrected *d'* for each test and the significance levels are depicted in the table on the bottom right.

($t = -1.43, p = .155$). A more detailed analysis of the hits in the X-Ray BDT was performed to examine how the rather large difference between hits and true hits came about. Figure 6.10 shows the different types of hits that could result. Total hits are the uncorrected hits, that is, those resulting from correctly judging a bag as NOT OK that actually contains a threat. With signal detection analysis, which was used for assessment of test results in certification until now, these total hits are taken into account. True hits are those hits where the candidates correctly marked the threat item, in this case the IED, in the image. False clicks are hits but where the candidates clicked on a wrong object in the bag. And false hits are hits where the candidates did not mark anything in the image. That is, false clicks and false hits are hits that are counted as hits when analyzing according to signal detection theory, that is, taking into account every hit that resulted from correctly judging a bag as NOT OK, whereas for the corrected analysis only true hits are taken into account. Figure 6.10 shows that for the X-Ray BDT candidates mostly marked some object in the image and therefore the difference between total hits and true hits are false clicks. False hits are only responsible for a fractional portion of the hits. Pearson

Test	r Hit rate - Proportion of true hits	p
ORT	0.071	0.033*
CAT	0.091	0.009*
BDT	-0.039	0.631

Table 6.4. Pearson correlation between the hit rate and the proportion of true hits for each test including the significance level.

correlations between the hit rate and the proportion of true hits on hits (#true hits / #hits) were calculated to examine if those candidates with more hits also achieved more true hits. Correlation results can be seen in Table 6.4. The results in Table 6.4 indicate that this seems not to be the case. That is, if a candidate achieves a high true hit rate does not depend significantly on the achieved hit rate.

Reliability

Cronbach's alpha and Guttman split half reliabilities were calculated for each test based on percent correct for Noise and Signal-plus-Noise trials separately. For the

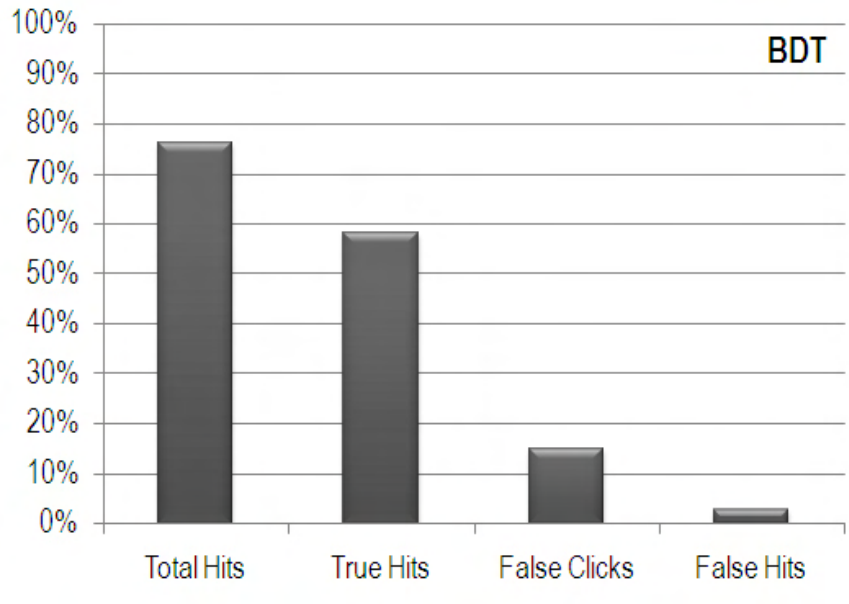


Fig. 6.10. Detailed graph of the hits for X-Ray BDT, averaged over all candidates and separated for the three types of hits: true hits are correctly identified, i.e. marked IEDs; false clicks are lucky hits because a wrong object was identified; false hits are lucky hits because nothing was marked in the bag. True hits, false clicks, and false hits sum up to total hits.

Signal-plus-Noise trials two different analyses resulted. Reliability was once calculated with percent correct according to the true hit system, that is, only correct rejections and true hits were coded as correct (SN - Click Analysis). And once reliability was calculated with percent correct according to signal detection analysis, that is, also false hits counted as correct (SN - Signal Detection). Noise trials are not affected by the different coding since it pertains only to the counting of hits and therefore only to Signal-plus-Noise images. Table 6.5 displays the reliability values.

6.4.3 Discussion

This study confirmed the expectations that in X-ray image interpretation tests, where images of bags have to be judged whether they contain a prohibited item or not, not an inconsiderable amount of hits (correctly judged bags as containing a prohibited item) result from a in fact harmless object. That is, the bags are very well classified as dangerous in that they contain prohibited items, however the actual prohibited items were not detected. Instead, harmless objects are interpreted as

Table 6.5. Reliability analyses for the X-Ray ORT, X-Ray CAT, and X-Ray BDT. Cronbach's alpha and Gutman split half values for Signal-plus-Noise trials (SN) and Noise trials (N) separately. Click Analysis shows the values based on percent correct according to true hits. Signal Detection shows the values based on percent correct according to signal detection theory.

	SN - Click Analysis		SN - Signal Detection		N	
	Alpha	Split half	Alpha	Split half	Alpha	Split half
ORT	.924	.871	.858	.745	.900	.851
CAT	.927	.903	.918	.888	.933	.928
BDT	.926	.848	.950	.877	.966	.917

being prohibited. As the results indicate, the hit rate as well as the *d'* scores decrease when only true hits are taken into account. The most striking decrease is to be found for the X-Ray BDT. Here, the decrease is almost 18%. If the uncorrected hit rate or the uncorrected *d'* scores are considered the mean test score is at the same level as for the X-Ray CAT. However, when only the true hits are included the performance decreases substantially (whereas the decreases for the X-Ray ORT and the X-Ray CAT are much smaller and quasi negligible with less than 5%). It seems as if in the X-Ray BDT the candidates are not able to detect some of the improvised explosive devices but still assess the bags as dangerous. This is all the more interesting because the false alarm rate is equal for the X-Ray BDT and the X-Ray CAT (this can be deducted from the *d'* value which is nearly the same for both tests and can also be confirmed by the data. If *d'* and the hit rate are comparable the false alarm rate must be as well). This implies that the candidates must have some "feeling" that an improvised explosive is in the bag though without actually detecting it. This can be confirmed by the data displayed in Figure 6.10, which show that the portion of hits that are not true hits, are actually false clicks. This means, candidates marked some suspicious object in the bag, not hitting the actual IED though. Since candidates know that in this test only IEDs have to be detected, the probability that they "detected" a threat object of another category should be rather small. The large standard deviations for the X-Ray BDT are replicating findings of earlier studies which also found a large increase in IED detection after training, even larger

than for all other threat categories, and a generally larger standard deviation for IED detection than for all other threat categories (Koller et al., 2008; Schwaninger & Hofer, 2004; Schwaninger, Hofer, & Wetter, 2007). These facts indicate large differences in the performance of IED detection between the candidates. Schwaninger and Hofer (2004) and Schwaninger, Hofer, and Wetter (2007) could also show that the performance increase for IED detection depends on the amount of training in that more training leads to a larger performance increase. So the large standard deviations for the X-Ray BDT could have their seeds in the fact that detection performance of IEDs and its increase is heavily dependent on the amount of training which in turn can differ remarkably between the candidates, as does the capability of IED detection. The correlations between the hit rate and the proportion of the true hits on the hits show that there is no evidence that higher scoring candidates - regarding hits - are generating more true hits, in relation to their hits, than lower scoring candidates. Correlations for the X-Ray CAT and the X-Ray ORT are significant; however, in consideration of the small effect size the effect can be neglected (Cohen, 1988). For the IED detection this means that training does not help to increase the proportion of true hits although the detection performance, when calculated with uncorrected values, can be increased significantly. This leads to the assumption that training mainly reduces the false alarm rate at this stage.

The correlations between the hit rate and the true hit rate and between the uncorrected *d'* and the corrected *d'*, respectively, are large and significant. This indicates that the same thing is measured with the two ways and therefore both ways of analysis could be used. However, the reliability has to be taken into account. For the X-Ray ORT and the X-Ray CAT the calculation based on true hits yields higher correlation coefficients, for Cronbach's alpha as well as for split-half. But for the X-Ray BDT, true hit analysis yields lower reliabilities than signal detection analysis. Interestingly, this is for the test with the largest difference between hits and true hits. Although the difference is rather small and the correlations between hit and true hit rate and between uncorrected and corrected *d'* are large and significant, it seems that for an IED detection test signal detection analysis would be appropriate to apply. Since the correlation between the hit rate and the proportion of

true hits is not significant for the X-Ray BDT, the data seems to indicate that even with training - assuming that candidates achieving a higher hit rate performed more training - actual detection and identification of IEDs is not given for a larger portion of the hits than with little training. Indeed, the percentage of true hits does not change significantly. However, besides reliability also the practical meaning should be considered.

The discussion about the analysis of test results mainly affects evaluation of certification. It apparently makes a difference in the evaluation if all hits or only true hits are counted. As mentioned earlier, in the work setting it is mostly kind of an irrelevant matter of fact how a hit comes about. The suspicious bag would be opened and searched and the threat item should then be detected. However, it has to be taken into account that some items should not be interpreted wrongly. For example, opening a bag based on the suspicion of finding a knife and instead finding an IED would not be in accordance with the rules and regulations. This argument supports also the definition of false hits and false clicks as misses. However, a basic decision should be taken about the importance of true hits and whether to focus on the assessment of real detection and identification competence or rather on effectiveness.

6.5 General discussion

The two experiments of this study showed that it is worthwhile taking a closer look at the analysis of tests. In Experiment 1 we could show that, depending on the scoring method, test scores in a multiple-true-false test could be increased. The lowest test scores were obtained with the MR method and the highest ones with the MTF scoring. With the MR method a point is awarded only if all response options of an item are marked correctly in contrast to the MTF scoring, where all true-false options which were answered correctly are credited evenly. Our findings are consistent with the ones of Albanese and Sabers (1988). In short, to assess the theoretical knowledge using a multiple choice test depending on the area, it has to be decided if partial knowledge shall be awarded or not. From our point of view it makes sense to award partial knowledge in the field of aviation security. However,

it has to be considered if specific questions do not allow for partial knowledge and should be analyzed differently. Contrary to Albanese and Sabers (1988) we found no difference in the reliability using different scoring methods. Further, the influence of the instruction on the test scores and reliability was tested. The group which was informed whether a question had one or more correct options achieved higher test scores. However, this specific information led to remarkably lower reliability values. It can be assumed that high performers can solve an item correctly independent of the information. Thus, no information differentiates more across the test candidates which in turn leads to higher reliability values. In Experiment 2 the influence of the analysis method on test scores and reliability measures in different image interpretation tests (X-ray screening tests) was investigated. Results showed lower hit rates and *d'* values if aviation security screening officers had to mark the correct prohibited item instead of only clicking on the NOT OK button. This decrease was highest in the X-Ray BDT. Further, reliability values changed differently. For the X-Ray ORT and X-Ray CAT reliability values were higher using the true hits instead of the total hits. Contrary results were found for the X-Ray BDT. However, it has to be mentioned that all reliability values are rather high and the changes rather small. To summarize, different ways of analyzing a test exist and mostly they generate different results, be it in terms of mean scores or in terms of reliability scores they yield. While a different mean score can be taken account of by adjusting the pass mark which defines if a candidate passes or fails the test, reliability is an innate feature of a test. Therefore, it is suggested that a test should be analyzed with the method that generates the best reliability. Furthermore, the analysis can focus on different aspects of a task. For example, in the case of X-ray image tests for certification of aviation security screening officers there has to be taken the decision if candidates are assessed regarding their real detection and identification performance for threat objects or regarding the efficiency their test behavior would imply at the checkpoint. For multiple choice tests, the different scoring methods should be taken into consideration and it should be decided whether partial knowledge should be awarded or not. As mentioned earlier, reliability values should always be taken

into account as only tests with high reliability allow for a fair measurement of the individual knowledge.

7

Applying Angoff Methods in Aviation Security X-Ray Screening: Can Angoff Methods be Applied to Airport Security Certification Tests?

7.1 Introduction

There are two different approaches in defining a standard (i.e., a pass mark) for a test. The norm-referenced approach is based on already obtained data. That is, the pass mark is established on the basis of test results obtained by the candidates. It is decided, for example, that a maximum of 300 candidates can pass the test (e.g., for an intermediate testing at the University, where the number of allowed students is restricted). This means that the pass mark is the score which lies between the test result of the best 300 candidates and the test result of the remaining candidates. Another means of setting the pass mark can be to take other statistical measures, for example the mean score minus one standard deviation. In a norm-referenced approach the pass mark varies depending on the sample. If a candidate passes or fails the test depends on the norm, that is, the sample he belongs to. This approach implies that the pass mark cannot be set before applying the test, giving it a slightly random touch. On the other hand, institutions applying the test can react to situational conditions. For example at the airport, a certain number of screeners have to pass certification in order to adhere operations. This can be governed by setting the pass mark accordingly so that, for example, only a certain number of candidates fail the test. Furthermore, the effort of establishing the pass mark is minimal. The criterion-referenced approach solves the problem of the normative approach (i.e., that the pass mark can only be set after application of the test) by setting a pass mark without normative data. Here, a pass mark is defined per test on the basis of

the difficulty of the individual items of the test. If a candidate scores higher than this pass mark, he or she passes the test, independently of how the other candidates in the sample score. One of the criterion-referenced methods is the Angoff method (Angoff, 1971). According to the *Criterion referenced performance standard setting* (2004), the level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation and should not be adjusted to regulate the number or proportion of persons passing the test. The pass point should be determined by careful analysis and judgment of acceptable performance. Due to the relatively simple process of determining the pass points the Angoff method is probably the most basic form of the criterion-based standard setting (Khalid & Saeed, 2007). The Angoff method includes the definition of a "minimally competent candidate", which is usually made in discussion with the judges. The judges usually are experts in the field of the test topic. Judges in this method are expected to review each test item and a passing score is computed from an estimate of the probability of a minimally acceptable candidate answering each item correctly. After discussion and consensus of the characteristics of a minimally acceptable candidate, each judge makes an independent assessment of the probability for each item that a minimally acceptable candidate will answer the item correctly. To determine the probability of a correct response for each item, the judges' assessments of the items are averaged. Then, these probabilities for all items of the test are averaged to obtain the pass point (*Criterion referenced performance standard setting*, 2004). The Angoff method features several advantages, that is, it is easy to implement, understand, and compute (Berk, 1986). Besides these advantages, there are also disadvantages of the Angoff method. First, it assumes that judges have a good understanding of the statistical concepts, and second, the panelists may lose sight of the students' overall performance on the assessment due to the focus on individual items, as this method carries an item-based procedure (Khalid & Saeed, 2007). Furthermore, the continuum of item probabilities tends to result in considerable variability among judges. Many judges have difficulty, defining students who are minimally competent (Berk, 1986). The Angoff method has undergone many alterations (Berk, 1986). Four of these alterations are

used in this study: the original Angoff method with and without discussion and the two-choice Angoff method with and without discussion. The original Angoff method as proposed by Angoff (1971) requires a number of judges to estimate with which probability a minimally competent candidate would solve each item of a test. In a slight modification of the original Angoff method, Angoff (1971) proposes that the judges think of how many of 100 minimally competent candidates would solve the items instead of the probability of one candidate to solve the items. It is assumed that this approach is easier to imagine. The first alteration is called two-choice Angoff method (Berk, 1986). The modification here consists of restricting the possible number of answers. The judges just have to say if a minimally competent candidate will or will not solve an item instead of giving the probability with which this candidate will solve the item. Both these approaches can also be applied with discussion. The judges discuss their difficulty estimates with each other after having given them separately first. After the discussion the judges can alter their estimates if they want. These alterations are called Iterative Angoff method (Berk, 1986) and Iterative two-choice Angoff method, respectively (Berk, 1986; Cross, Impara, Frary, & Jaeger, 1984; Jaeger, 1982). In both alterations there is the possibility to provide the judges with normative data for their discussion. These four versions have been studied many times regarding written, especially multiple-choice, tests. In this experiment we want to examine if these methods can be applied to X-ray image tests and if yes, which one is the best, regarding reliability, validity and convenience. X-ray image tests can be used for certification of aviation security screening officers and consist of X-ray images of passenger bags which have to be judged regarding dangerousness of their content. That is, the test items are two-alternative forced choice items with the responses dangerous (NOT OK) and not dangerous (OK), respectively.

7.2 Method

7.2.1 Participants

Twenty-two (seven men) aviation security screeners participated in this study as judges. They were randomly assigned to one of two groups. All of them were selected based on their detection performance in X-ray screening. To this end, certification results and TIP[1] data were taken into account. All screeners had at least three years experience on the job.

7.2.2 Materials

To test all four Angoff methods, the X-Ray Competency Assessment Test (X-Ray CAT) by Koller and Schwaninger (2006) was used. This X-ray image test consists of 256 X-ray images of passenger bags whereof 128 images include a prohibited item (e.g., a gun, knife, improvised explosive device (IED), or other threat item like grenade, etc.). In the test itself images are displayed for 15 seconds on the screen. Then participants have to decide whether the bag was OK or NOT OK regarding allowance to be transported on board. Additionally, they have to indicate how sure they are in their decision by clicking on a 50 point slider. For all bags which were judged as NOT OK they have to indicate to which of the four categories (gun, knife, IED, other) the threat item belongs to. For the first part of the study in which participants have to judge each image of the X-Ray CAT we used the computer program PCQuest. In the second part all X-ray images were projected and printed out for discussion. The X-Ray CAT was used for certification of the aviation security screeners in 2007.

7.2.3 Procedure

Both Angoff methods (original and two-choice) were explained. A minimally competent candidate was not defined; instead the participants were instructed to focus

[1] Threat Image Projection (TIP) automatically inserts images of fictitious threat items into the X-ray image of passenger bags at the security checkpoint at an airport. TIP performance takes into account how many of these virtually inserted threat objects are detected by the screening officer on duty.

on the performance of two or three minimally competent aviation security screeners they know, that is, those who would just be sufficiently capable of doing their job. They shortly discussed their task and these "borderline" screeners. Participants were informed that after their judgment of all items there will be a discussion and that the goal will be to talk the differences in their judgments over. Before participants started they received a short introduction and some exercise trials to familiarize with the task. Within the first part, participants had the possibility to take a short break if desired. The first part of the experiment was conducted computer-based. Before participants judged each image regarding its difficulty to be solved correctly they had to indicate by themselves whether the bag was OK or NOT OK and how sure they are in their decision (same as in the test setting). Then they saw the image again with the prohibited item depicted separately in case the X-ray image included a prohibited item. One group had to indicate whether a minimally competent screener would judge the X-ray image correctly or not (two-choice Angoff method, see Figure 7.1), the other group had to indicate with which probability a minimally competent screener would judge the X-ray image correctly (classical Angoff method, see Figure 7.2). The latter group was instructed that they should not take into account guessing probability (which in this case would be 50%), but use the whole scale (0-100%) for the difficulty estimation.

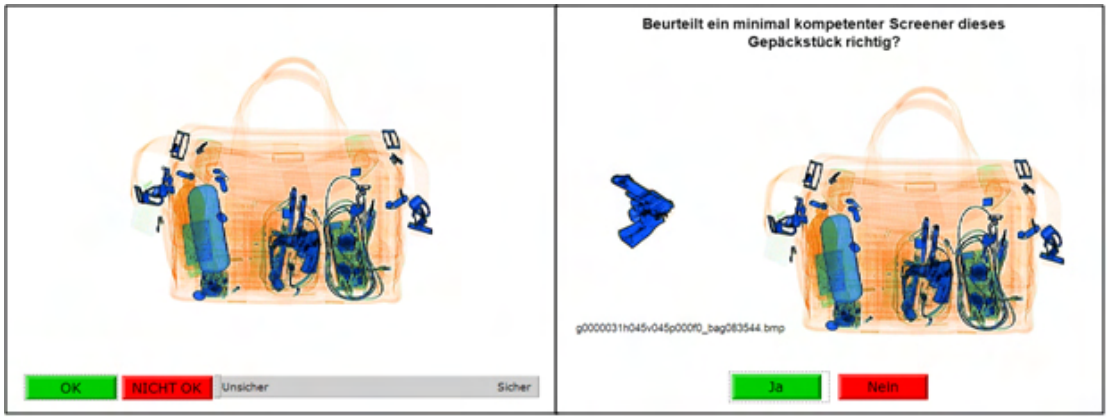


Fig. 7.1. Screenshot of the experimental interface applying the two-choice Angoff method. In a first step participants had to judge each item (left). In a second step the had to decide if a minimally competent candidate would solve the item or not (right). In case the image contains a threat object it was depicted separately.

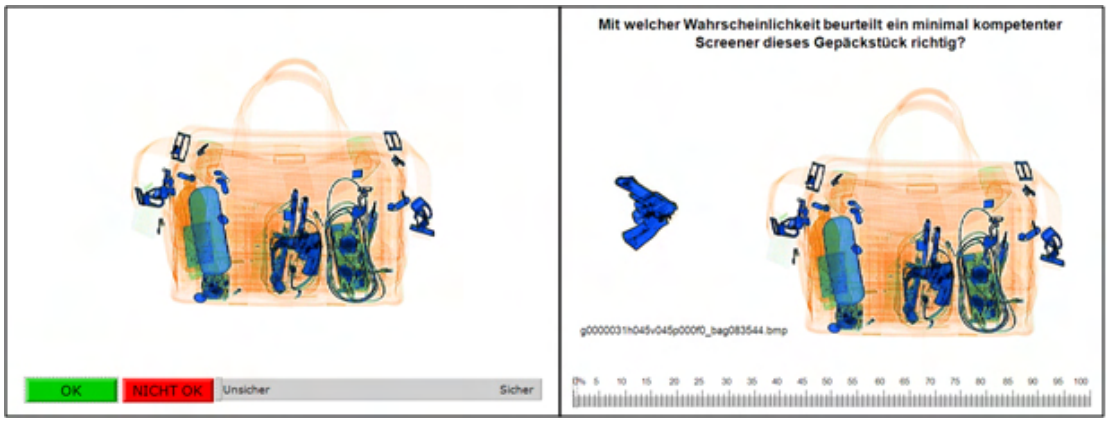


Fig. 7.2. Screenshot of the experimental interface applying the classical Angoff method. In a first step participants had to judge each item (left). In a second step the had to decide with which probability a minimally competent candidate would solve the item (right). In case the image contains a threat object it was depicted separately.

After completing the first sequence participants had to relax for at least 30 minutes. After this break the discussion was started with both groups separately. In this part participants had to discuss their decisions on each item within their group. To this end, participants had a look at every test item once again. Pictures were presented in random order. Participants had to discuss those items where the standard deviation of their estimates was bigger than the average of the deviations plus one standard deviation ($_{SD}+SD_{SD} < SD$). But they could discuss every other item as well. During the discussion participants were allowed to change their estimates if they wished so but were not forced to do so. Any changes were recorded. Discussions lasted about 2.5 (classical Angoff method) and 3 (two-choice Angoff method) hours. Neither of the two groups discussed every item.

7.3 Calculations

First of all, it is important to assess whether the ratings obtained are reliable, that is, whether there is consistency among raters or not. We have decided to apply two measures to assess reliability of the ratings, namely Cronbach's Alpha and intraclass correlations. Cronbach's Alpha is a widely recognized measure for the internal consistency of a scale based on the average inter-item correlation. In this case, Cronbach's

Alpha as the interrater correlation was calculated as measure of internal consistency between the ratings of the different judges. The formula of Cronbach's Alpha is

$$\alpha = (N \times \overline{r}/(1 + (N - 1) \times \overline{r})$$

whereas N equals the number of participants and $\overline{r}$ is the average interitem correlation among the participants, respectively. Intraclass correlation coefficients assess interrater reliability. A "two-way mixed model for absolute agreement" was chosen since raters were deliberately selected and since the aim is to have absolute (not only relative) agreement of the raters. In a next step, the hit rate (pHit) from the threat images ("signal plus noise", SN, i.e., those bags containing a threat object) and the false alarm rate (pFA) from the non threat images ("noise", N, i.e., those bags not containing a threat object) were calculated. (Note that the false alarm rates for the non threat images were obtained indirectly by subtracting the rating value from one, since the raters assessed the probability of a correct answer (i.e., correct rejection) and not a false alarm.) Since it is used in many scientific publications regarding X-ray image interpretation competency and during training and testing procedures we chose to compute the non-parametric performance measure A' which is based on pHit and pFA (Pollack & Norman, 1964; Macmillan & Creelman, 1991; see Hofer & Schwaninger, 2004, for the application of detection theories to X-ray screening), using the following formula (Grier, 1971):

$$A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)]$$

whereas H is the hit rate and F the false alarm rate. If the false alarm rate is greater than the hit rate, the equation must be modified (Aaronson & Watts, 1987; Snodgrass & Corwin, 1988):

$$A' = 0.5 - [(F - H)(1 + F - H)]/[4F(1 - H)]$$

The term non-parametric refers to the fact, that the computation of A' requires no a priori assumption about underlying distributions. A' can be calculated when the validity of the normal distribution and equal variance assumptions of the signal-noise and noise distribution cannot be verified.

7.4 Results

7.4.1 Reliability

Cronbach's Alpha (see above for details on the calculation) and intraclass correlations were applied on hit rates and false alarm rates. As shown in Tables 7.1 and 7.2, Cronbach's Alpha was calculated separately for all four methods. Further, scores were calculated for all images together, as well as for N and SN images only.

Table 7.1. Cronbach's alpha for the ratings with the classical Angoff method

Classical Angoff	All images	SN only	N only
Without discussion	.850	.894	.768
After discussion	.949	.946	.952

Table 7.2. Cronbach's alpha for the ratings with the two-choice Angoff method

Two-choice Angoff	All images	SN only	N only
Without discussion	.881	.951	.821
After discussion	.952	.962	.937

Similarly, intraclass correlation coefficients were calculated separately for all four methods. It was again distinguished between all images together, all threat images (SN) separately, and all non threat images (N) separately, as shown in Tables 7.3 and 7.4.

Table 7.3. Intraclass correlation coefficients for the ratings with classical Angoff method

Classical Angoff	All images	SN only	N only
Without discussion	.268	.325	.176
After discussion	.591	.574	.609

Table 7.4. Intraclass correlation coefficients for the ratings with the two-choice Angoff method

Two-choice Angoff	All images	SN only	N only
Without discussion	.384	.481	.258
After discussion	.641	.694	.566

7.4.2 Criterion

As a measure of detection performance A' was calculated for the X-Ray CAT (see above for details on the calculation). A' for the classical Angoff method is .833 (SD = 0.067). For the iterative Angoff method, it is .847 (SD = 0.056), respectively. A' for the two-choice Angoff method is .811 (SD = 0.098). For iterative, two-choice Angoff method it is .838 (SD = 0.035), respectively. The mean A' for all certified screeners (n = 764) in 2007 was .898 (SD = .042). With the normative pass mark, set by the Swiss Federal Office of Civil Aviation, 5 screeners (1%) failed the test (*Note*: the pass mark cannot be published for security reasons). When applying the criterion-referenced pass marks the failing rate is as follows: with classical Angoff method 65 screeners (9%) would fail, with iterative Angoff method 86 screeners (11%) would fail, with two-choice Angoff method 31 screeners (4%) would fail, and with iterative two-choice Angoff method 65 screeners (9%) would fail the X-Ray CAT certification test.

7.4.3 Correlations

Table 7.5 shows all calculated Spearman correlations. We correlated the average rating per image assessed by the raters for all four methods with each other and with the hit rate (for threat images) and the correct rejection rate (for non threat images) from the X-Ray CAT certification test 2007. Significant correlations are depicted in scatter plots in Figures 7.3-7.8.

Table 7.5: Spearman correlations and p-values of average ratings and CAT certification data

	All images	SN only	N only
Average classical Angoff method - Average iterative Angoff method	$r = -.030$ $p = .631$	$r = -.034$ $p = .704$	$r = -.047$ $p = .598$
Average two-choice Angoff method - Average iterative two-choice Angoff method	$r = .821$ $p < .001$	$r = .879$ $p < .001$	$r = .739$ $p < .001$

Continued on Next Page...

Table 7.5: Spearman correlations and p-values of average ratings and CAT certification data

	All images	SN only	N only
Average classical Angoff method -	$r = .793$	$r = .834$	$r = .739$
Average two-choice Angoff method	$p < .001$	$p < .001$	$p < .001$
Average iterative Angoff method -	$r = .004$	$r = -.077$	$r = .084$
Average iterative two-choice Angoff method	$p = .943$	$p = .385$	$p = .644$
Average classical Angoff method -	$r = .719$	$r = .802$	$r = .608$
Average iterative two-choice Angoff method	$p < .001$	$p < .001$	$p < .001$
Average iterative Angoff method -	$r = -.099$	$r = -.141$	$r = -.056$
Average two-choice Angoff method	$p = .114$	$p = .112$	$p = .344$
Average classical Angoff method -	$r = .740$	$r = .847$	$r = .602$
CAT certification test	$p < .001$	$p < .001$	$p < .001$
Average iterative Angoff method -	$r = -.035$	$r = -.043$	$r = -.077$
CAT certification test	$p = .575$	$p = .644$	$p = .390$
Average two-choice Angoff method -	$r = .712$	$r = .818$	$r = .569$
CAT certification test	$p < .001$	$p < .001$	$p < .001$
Average iterative two-choice Angoff method -	$r = .643$	$r = .775$	$r = .472$
CAT certification test	$p < .001$	$p < .001$	$p < .001$

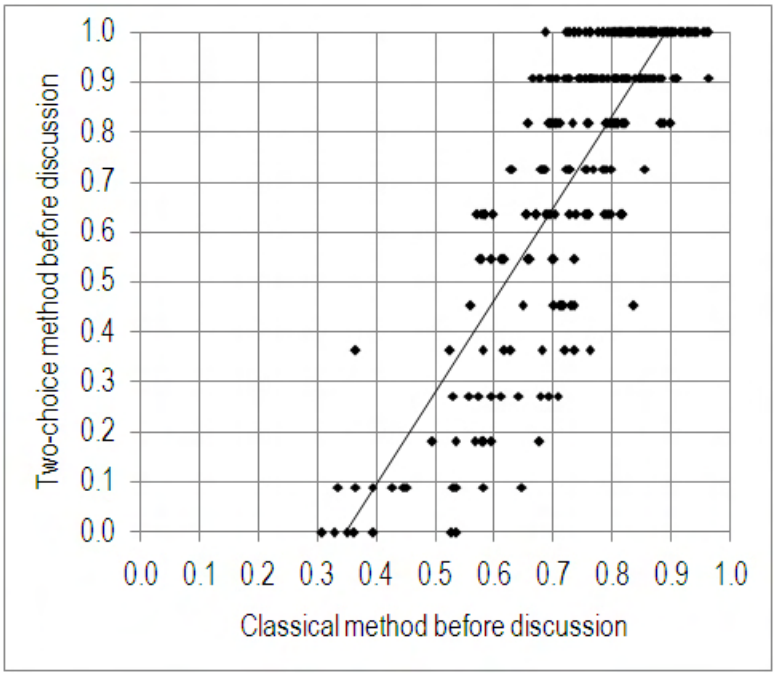


Fig. 7.3. Scatter plot of judges' averaged estimates of the probability that a minimally competent candidate will solve a test item, by two-choice Angoff method before discussion against classical Angoff method before discussion. Each dot represents a test item.

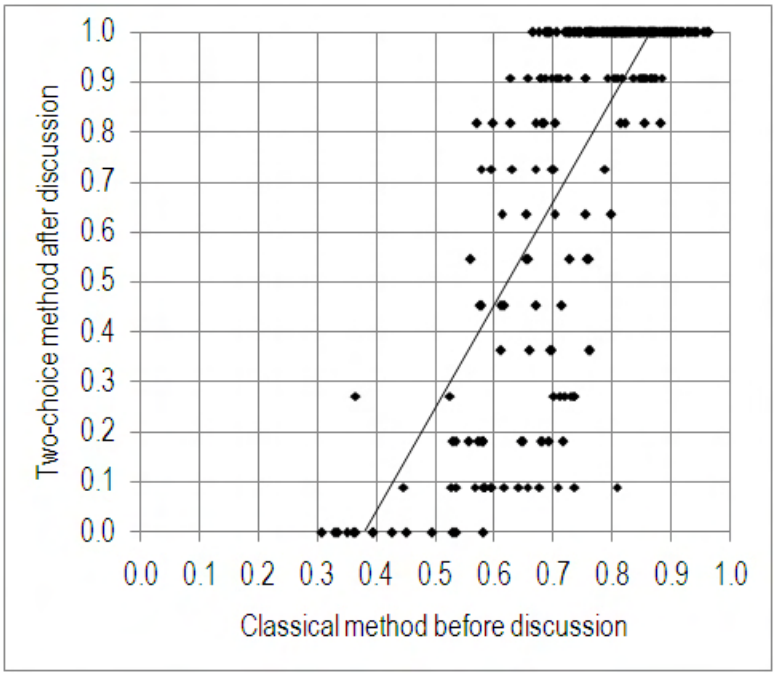


Fig. 7.4. Scatter plot of judges' averaged estimates of the probability that a minimally competent candidate will solve a test item, by two-choice Angoff method after discussion against classical Angoff method before discussion. Each dot represents a test item.

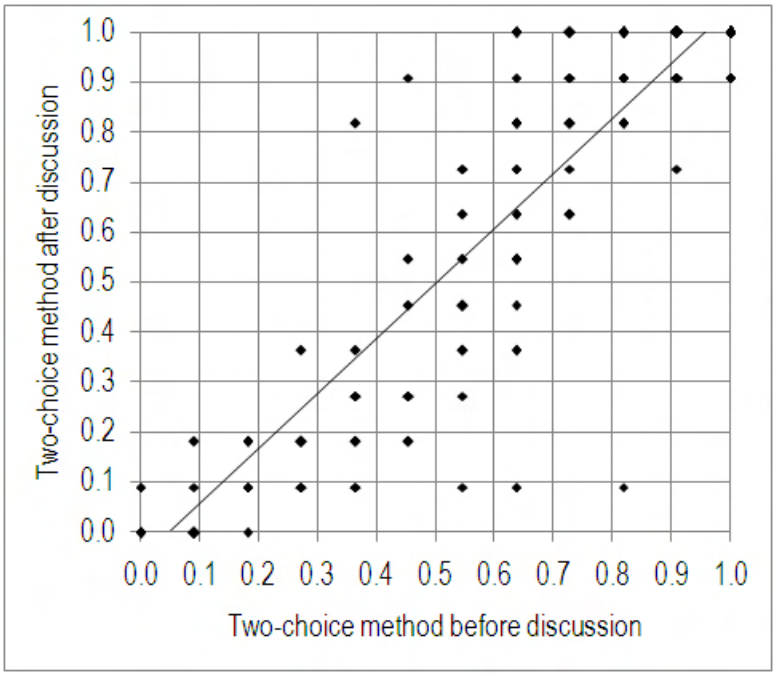


Fig. 7.5. Scatter plot of judges' averaged estimates of the probability that a minimally competent candidate will solve a test item, by two-choice Angoff method after discussion against two-choice Angoff method before discussion. Each dot represents a test item. Bigger dots comprise more than one item.

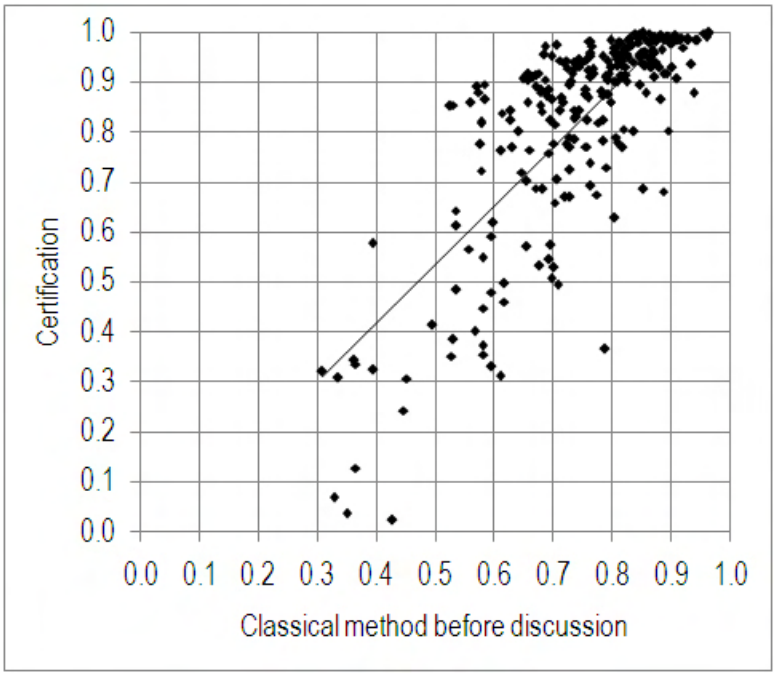


Fig. 7.6. Scatter plot of actual performance values for each test item in the certification against judges' averaged estimates of the probability that a minimally competent candidate will solve a test item by classical Angoff method before discussion. Each dot represents a test item.

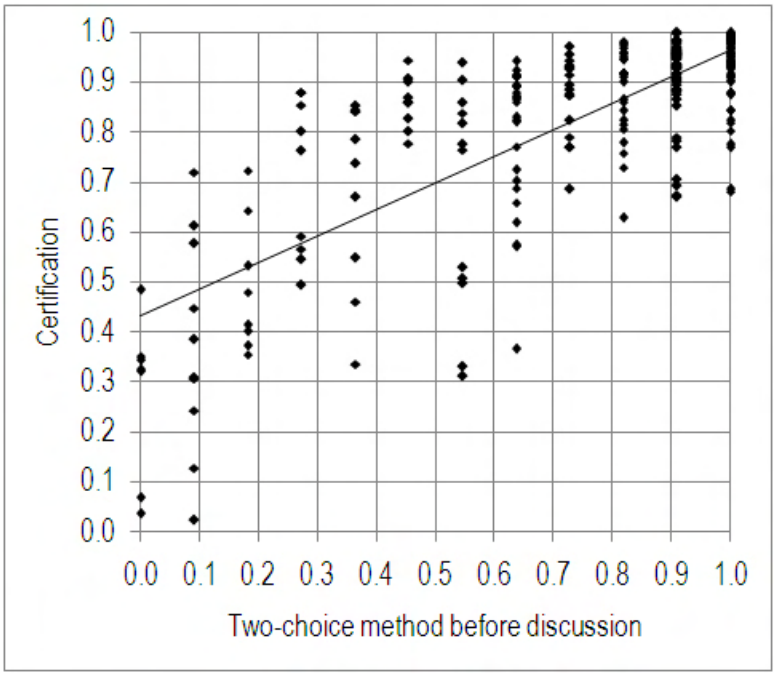


Fig. 7.7. Scatter plot of actual performance values for each test item in the certification against judges' averaged estimates of the probability that a minimally competent candidate will solve a test item by two-choice Angoff method before discussion. Each dot represents a test item.

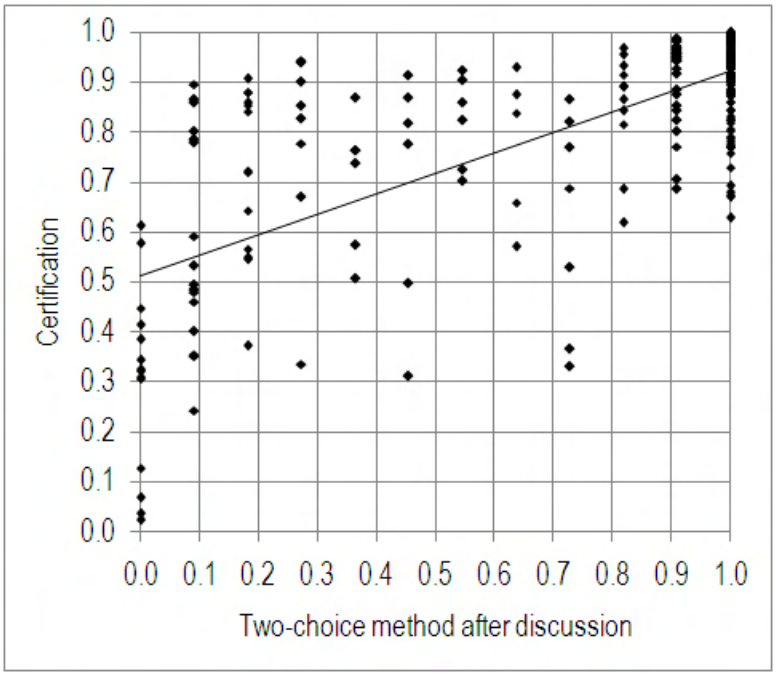


Fig. 7.8. Scatter plot of actual performance values for each test item in the certification against judges' averaged estimates of the probability that a minimally competent candidate will solve a test item by two-choice Angoff method after discussion. Each dot represents a test item.

Figure 7.9 shows the scatter plot of the difficulty estimates in the classical Angoff method before against after discussion.

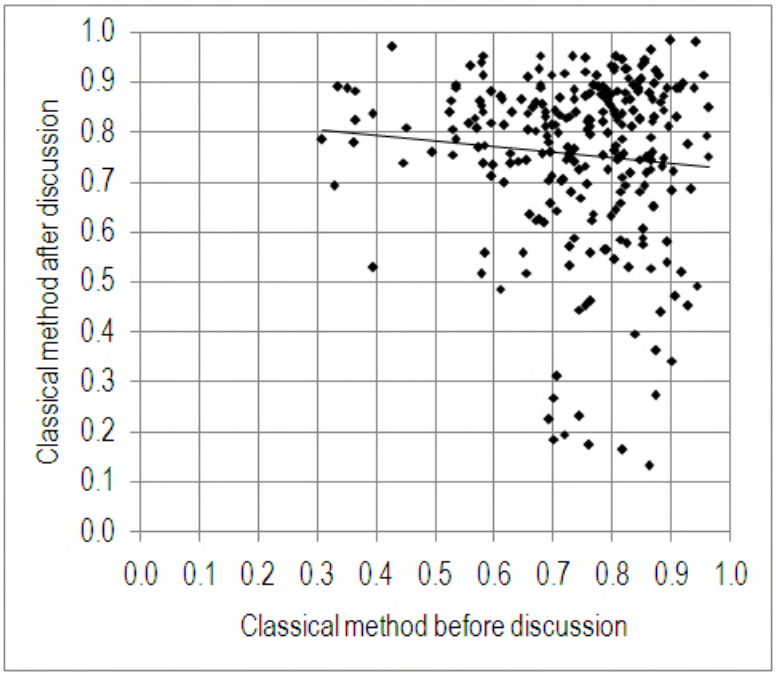


Fig. 7.9. Scatter plot of judges' averaged estimates of the probability that a minimally competent candidate will solve a test item by classical Angoff method after discussion against classical Angoff method before discussion. Each dot represents a test item.

7.5 Discussion

Both the classical as well as the two-choice Angoff method resulted in good reliabilities (see Tables 7.1 and 7.2). All reliability values are well above .80, except for the N images in classical Angoff method without discussion. Furthermore, the reliabilities are in a comparable range for all methods with slightly higher coefficients for the iterative methods (with discussion). This is all the more the case for the intra-class correlations (see Tables 7.3 and 7.4), where correlation coefficients before discussion are rather low and increase substantially after discussion. The natural reason for this is that after discussion judges mostly harmonize their difficulty estimates of the items and therefore smooth the discrepancies between themselves out. This way, the agreement of the raters and thus the interrater reliability increases. Since the Cronbach's alpha reliabilities are in the same range before and after discussion, the discussion for setting a standard can be left out as this would save much time. When comparing the mean A' detection performance values as obtained with the four different methods there are only marginal differences. This indicates that on average all methods result in a comparable mean detection performance - and therefore pass mark - through the judgment of the single test items. However, if compared with

the normative pass mark set by the Swiss Federal Office of Civil Aviation, criterion-referenced pass marks as obtained through the four Angoff methods in this study are higher with the effect that more screeners would fail the certification test. In the decision about which pass mark to use the actual purpose of a certification would stand against practical reasons. The actual purpose of a certification is to evaluate the knowledge and skills of practitioners seeking a credential to ensure that employees meet a desired level of competence (Althouse, 2000; Shimberg, 1981). Practical reasons however sometimes call for a generous adaptation of the pass mark for the simple reason of allowing to maintain operations. This implies a careful analysis of the situation and the requirements before defining a standard setting method and a pass mark.

As shown in Table 7.5, the estimates resulting from the iterative Angoff method neither correlate with any other Angoff rating nor with the results of the certification test 2007. This could be because the participants in this group adjusted their estimates after the discussion as follows: For SN trials they normally corrected their estimates to lower values. And for N trials they raised their ratings in most cases. In doing this they raised the correct rejection rates and lowered the hit rates (see Figure 7.9). Consequently, this method cannot be applied to the CAT certification test 2007.

Significant correlations were found between the two-choice and iterative two-choice method, between the classical and the two-choice method, less important between the classical and iterative two-choice method, and, more important, between the CAT test results and classical, two-choice, and iterative two-choice method. The correlation between the item difficulty estimates obtained through two-choice and iterative two-choice method (see Figure 7.5 signifies that the judges did not change their estimates significantly after discussion unlike the judges of the classical Angoff method group. Prior to the discussion the two groups estimated the test items rather similarly as indicates the significant correlation between the classical and the two-choice method (see Figure 7.3). As a consequence of these two facts the correlation between the classical and the iterative two-choice method is significant

as well (see Figure 7.4). Interesting with regards to the normative approach are the correlations between the Angoff estimates and the actual test results achieved by over 900 aviation security screeners during certification. Except for the iterative method all correlations were significant (see Table 7.5 and Figures 7.6 - 7.8). This indicates that the judges estimated the difficulty of the test items rather similarly to the actual performance of the normative group. These results qualify both standard setting approaches as appropriate, the normative approach as well as the criterion-referenced approach. This means that there is no substantive reason in favor or against one particular approach regarding statistical measures and the decision on which method to use to define a pass mark will have to be based on the advantages and disadvantages of the two approaches and be taken by each institution individually. While the effort of setting the pass mark is far smaller with the normative approach than with the criterion-referenced approach and the normative approach allows for adaptation of the pass mark to situational circumstances, the psychological effect may favor the criterion-referenced approach. Candidates may accept a criterion-referenced pass mark more easily than a norm-based. They know, if they achieve a certain minimum score they will pass the test independently of the other candidates' scores. This may seem to be fairer. With a criterion-referenced pass mark all candidates have the chance to pass the test. However, there is also the chance that all candidates or a major portion fails the test. If the situation does not allow for more than a certain number of candidates to fail the test the pass mark would have to be adapted.

If a choice would have to be made on which criterion-referenced approach to choose in the situation of aviation security screener certification tests the results of this study favor the two-choice Angoff method. The iterative two-choice Angoff method has the highest reliability but the lowest correlation with the results of the CAT certification test and the classical Angoff method has the highest correlation but the lowest reliability. Therefore, we suggest using the two-choice Angoff method for standard setting of this kind of test. It has a high and significant correlation with the results of the certification test and a good reliability. Furthermore, as compared to the classical Angoff method it is easy to apply and understand for the judges.

However, as mentioned earlier, the reliability differences between the four methods are rather small and in this case can primarily used as a decision aid on which method to use but do not disqualify any method from being used for this purpose.

In summary, it can be said that the Angoff methods are appropriate to use for standard setting for X-ray image tests. The results obtained in this study indicate that all methods would result in similar pass marks. Which method should be the first choice - whether it is norm-based or criterion-referenced - depends on the purpose of the test and the situation.

References

Aaronson, D., & Watts, B. (1987). Extensions of grier's computational formulas for a' and b" to below-chance performance. *Psychological Bulletin*, *102*, 439–442.

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1979). Clueing in multiple-choice test items with combinations of correct responses. *Journal of Medical Education*, *54*(12), 948-950.

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false item: a study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement*, *25*(2), 111-123.

Althouse, L. A. (2000). Test development: Ten steps to a valid and reliable certification exam. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, April 9-12.* Indianapolis, Indiana.

Angoff, W. H. (1971). Norms, scales, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington: American Council on Education.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement*, *26*, 323-336.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying mutliple choice tests: A case study in irrationality. *Mind and Society*, *4*, 3-12.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137–172.

Bernstein, A., & Li, J. (2005). *From active towards interactive learning: Using consideration information to bias error-prone human experts into correct labeling.* (Unpublished manuscript)

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115-147.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 60-64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *3*, 247-260.

Brandenburg, D., & Whitney, D. (1972). Matched pair true-false scoring: Effect on reliability and validity. *Journal of Educational Measurement*, *9*(4), 297-302.

Buckley-Sharp, M., & Harris, F. (1971). The scoring of multiple-choice questions. *British Journal of Medical Education*, *5*, 279-288.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, *19*, 1-19.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement*, *32*, 1035-1038.

Costin, F. (1976). Difficulty and homogeneity of three-choice versus four-choice objective test items when matched for content of stem. *Teaching of Psychology*, *3*, 144-145.

Craik, F. I., & Salthouse, T. A. (2000). *The handbook of aging and cognition.* Mahwah, NJ: Erlbaum.

Criterion referenced performance standard setting. (2004). http://www.measurementresearch.com/wwa/default.shtml.

Cross, L. H., & Frary, R. B. (1977). An empirical test of lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement*, *14*(4), 313-321.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the national teacher examinations. *Journal of Educational Measurement*, *21*(2), 113-129.

Cureton, E. E. (1966). The correction for guessing. *Journal of Experimental Education*, *34*(4), 44-47.

Czaja, S. J., & Drury, C. G. (1981a). Aging and pretraining in industrial inspection. *Human Factors*, *23*(4), 485-494.

Czaja, S. J., & Drury, C. G. (1981b). Training programs for inspection. *Human Factors I*, *23*(4), 473–484.

Davis, F. B. (1964). *Educational measurements and their interpretation.* Belmont, CA: Wadsworth.

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, *43*(2), 181-191.

Dollinger, S. M., & Hoyer, W. J. (1996). Age and skill differences in the processing demands of visual inspection. *Applied Cognitive Psychology*, *10*, 225–239.

Drury, C. G. (1975). Inspection of sheet metal materials: model and data. *Human Factors*, *17*, 257–265.

Drury, C. G., Ghylin, K. M., & Holness, K. (2006). Error analysis and threat magnitude for carry-on bag inspection. In *Proceedings of the 50th annual human factors and ergonomics society meeting, san francisco, ca* (Vol. 50, pp. 1189–1193).

Dugdale, A. E. (1971). A revised marking scheme for multiple-choice questions. *British Journal of Medical Education*, *5*, 162-164.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, *29*, 565-570.

Ebel, R. L. (1972). Why is a longer test usually a more reliable test? *Educational and Psychological Measurement*, *32*, 249-253.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, *32*, 2385-2400.

Fishman, J., & Galguera, T. (2003). *Introduction to test construction in the social and behavioural sciences. a practical guide.* Oxford: Rowman & Littlefield.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194-200.

Fozard, J. L. (1990). Vision and hearing in aging. In J. E. Birren, R. B. Sloane, & G. D. Cohen (Eds.), *Handbook of mental health and aging* (3rd ed., pp. 150–170). San Diego, CA: Academic Press.

Frary, R. B. (1969). Elimination of the guessing component of multiple-choice test scores: effect on reliability and validity. *Educational and Psychological Measurement*, *29*, 665-680.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, *4*, 79-90.

Frary, R. B. (1982). A simulation study of reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational Statistics*, *7*(4), 333-351.

Frary, R. B., & Zimmerman, D. W. (1970). Effect of variation in probability of guessing correctly on reliability of multiple-choice tests. *Educational and Psychological Measurement*, *30*, 595-605.

Frisbie, D. A. (1973). Multiple choice versus true-false: a comparison of reliabilities and concurrent validities. *Journal of Educational Measurement*, *10*(4), 297-304.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, *11*(4), 21-26.

Gale, A. G., Mugglestone, M., Purdy, K., & McClumpha, A. (2000). Is airport baggage inspection just another medical image? In *Medical imaging: Image perception and performance. progress in biomedical optics and imaging* (Vol. 1, pp. 184–192).

Gale, A. G., Purdy, K., & Wooding, D. (2005). Human factors in design, safety, and management. In D. de Waard, K. A. Brookhuis, R. van Egmond, & T. Boersema (Eds.), (p. 63). Shaker.

Ghylin, K. M., Drury, C. G., & Schwaninger, A. (2006). Two-component model of security inspection: application and findings. In *16th world congress of ergonomics, iea 2006, maastricht, the netherlands, july* (pp. 10–14).

Glass, G. V., & Wiley, D. E. (1964). Formula scoring and test reliability. *Journal of Educational Measurement*, *1*(1), 43-49.

Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2002). *Psychophysical results from experiments on recognition and categorisation.* Information Society Technologies (IST) programme, Cognitive Vision Systems - CogVis (IST-2000-29375).

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, *75*, 424–429.

Gronlund, N. E. (1965). *Measurement and evaluation in teaching.* New York, NY: Macmillan.

Gross, L. J. (1982). Scoring multiple true/false tests: Some considerations. *Evaluation and the Health Professions*, *5*(4), 459-468.

Gross, M. E., & Wright, B. A. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement*, *45*, 1-13.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*(1), 51-78.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-334.

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hamilton, R. (2003). Why do people suggest what they do not want? using context effects to influence others' choices. *Journal of Consumer Research*, *29*(4), 492-506.

Hammerton, M. (1965). The guessing correction in vocabulary tests. *British Journal of Educational Psychology*, *35*, 249-251.

Harasym, P., Norris, D., & Lorscheider, F. (1980). Evaluating student multiple-choice responses: Effects of coded and free formats. *Evaluation and the Health Professions*, *3*(1), 63-84.

Harden, R. M., Lever, R., & Wilson, G. M. (1969). Two systems of marking objective examination questions. *Lancet*, *1*, 40-42.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The object recognition test ort - a reliable tool for measuring visual abilities needed in x-ray screening. In *IEEE ICCST Proceedings* (Vol. 39, pp. 189–192).

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006a). Increased detection performance in airport security screening using the x-ray ort as pre-employment assessment tool. In *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28* (pp. 393–397).

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006b). The role of recurrent cbt for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. In *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 - December 1* (pp. 338–342).

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences*, *7*(10), 425-427.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 1511-1521.

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in x-ray screening. In *IEEE ICCST Proceedings* (Vol. 38, pp. 303–308).

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. In *WIT Transactions on the Built Environment* (Vol. 82, pp. 417–426).

Horst, P. (1933). The difficulty of multiple-choice test item. *Journal of Educational Psychology*, *24*, 229-232.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480-517.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: theory and application. *Educational Evaluation and Policy Analysis*, *4*(4), 461-475.

Jaradat, D., & Sawaged, S. (1986). The subset selection technique for multiple-choice tests: An empirical inquiry. *Journal of Educational Measurement*, *23*(4), 369-376.

Jolliffe, I. T. (2002). *Principle component analysis* (2nd ed.). New York: Springer-Verlag.

Khalid, M. N., & Saeed, M. (2007). Criterion referenced setting performance standards with an emphasis on angoff method. *Journal of Research and Reflections in Education*, *1*, 66–87.

Kline, P. (2000). *Handbook of psychological testing.* London: Routledge.

Klock, B. A. (2005). Test and evaluation report for x-ray detection of threats using different x-ray functions. In *IEEE ICCST Proceedings* (p. 182-184).

Koller, S., & Schwaninger, A. (2006). Assessing x-ray image interpretation competency of airport security screeners. In *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006* (p. 399-402).

Koller, S. M., Hardmeier, D., Michel, S., & Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent cbt of x-ray image interpretation. *Journal of Transportation Security*, *1*(2), 81–106.

Kolstad, R. K., Briggs, L. D., & Kolstad, R. A. (1985). Multiple-choice classroom achievement tests: Performance on items with five vs. three choices. *College Student Journal*, *19*, 427-431.

Kosslyn, S. M. (1994). *Image and brain. the resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kotler, P. (1988). *Marketing management: Analysis, planning, implementation and conrol* (6th ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on diferent item response formats - an experiment in fundamental research on psychological assessment. *Psychology Science*, *49*(4), 361-374.

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica*, *102*, 221-245.

Lawson, R., & Humphreys, G. W. (1996). View-specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 395-416.

Lawson, R., & Humphreys, G. W. (1998). View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. *Perception & Psychophysics*, *60*, 1052-1066.

Leach, J., & Morris, P. E. (1998). Cognitive factors in the close visual and magnetic particle inspection of welds underwater. *Human Factors*, *40*(2), 187-197.

Lennox, B. (1967). Marking multiple-choice examinations. *British Journal of Medical Education*, *1*, 203-211.

Little, E., & Creaser, J. (1966). Uncertain responses on multiple-choice examinations. *Psychological Reports*, *18*, 801-802.

Liu, X., Gale, A., Purdy, K., & Song, T. (2006). Is that a gun? the influence and features of bags and threat items on detection performance. In *Contemporary Ergonomics, Proceedings of the Ergonomic Society* (pp. 17–22).

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577-621.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*(1), 7-11.

Lowe, D. G. (1985). *Perceptual organization and visual recognition.* Boston: Kluwer Academic Publishing.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, *31*, 355-395.

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the "knowledge or random guessing" model. *Psychmetrika*, *69*(1), 147-159.

Macmillan, N. A., & Creelman, D. C. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Magnusson, D. (1967). *Test theory.* Reading, MA: Addison-Wesley.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. In *Proceedings of the Royal Society. Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 200, p. 269-294).

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. *Educational and Psychological Measurement*, *35*, 727-730.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport security screening. *Psychological Science*, *15*, 302–306.

McPhee, L. C., Scialfa, C. T., Dennis, W. M., Ho, G., & Caird, J. K. (2004). Age differences in visual search for traffic signs during a simulated conversation. *Human Factors*, *46*, 674-685.

Morawski, T., Drury, C. G., & Karwan, M. H. (1980). Predicting search performance for multiple targets. *Human Factors*, *22*, 707-718.

Morris, T. (2004). *Computer vision and image processing.* Basingstoke: Palgrave Macmillan.

Muijtjens, A. M. M., Mameren, H. van, Hoogenboom, R. J. I., Evers, J. L. H., & Vleuten, C. P. M. van der. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education*, *33*, 267-275.

Murphy, K. R., & Davidshofer, C. (2001). *Psychological testing.* Upper Saddle River, NJ: Prentice Hall.

Murray, F. S., & Szymczyk, J. M. (1978). Effects of distinctive features on recognition of incomplete pictures. *Developmental Psychology*, *14*(4), 356-362.

Murray, J. E. (1997). Flipping and spinning: Spatial transformation procedures in the identification of rotated natural objects. *Memory & Cognition*, *25*, 96-105.

Murray, J. E. (1999). Orientation-specific effects in picture matching and naming. *Memory & Cognition*, *27*, 878-889.

Newell, F. N., & Findlay, J. M. (1997). The effect of depth rotation on object identification. *Perception*, *26*, 1231-1257.

Owsley, C., Sekuler, R., & Siemsen, D. (1983). Contrast sensitivity throughout adulthood. *Vision Research*, *23*(7), 689-700.

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance ix* (pp. 135–152). Hillsdale, NJ: Erlbaum.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Review Neuroscience*, *5*, 291-303.

Peissig, J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 years ago? *Annual Review of Psychology*, *58*, 75-96.

Plude, D. J., & Hoyer, W. J. (1986). Age and the selectivity of visual information processing. *Psychology and Aging*, *1*(1), 4-10.

Plumlee, L. B. (1952). The effect of difficulty and chance success on item-test correlations and on test reliability. *Psychmetrika*, *17*, 69-86.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263-266.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*(1), 125–126.

Riegelnig, J., & Schwaninger, A. (2006). The influence of age and gender on detection performance and the criterion in x-ray screening. In *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006.*

Roberts, A. O. H. (1962). The maximum reliability of a multiple-choice test. *Psychologia Africana*, *9*, 286-293.

Roberts, J. H., & Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, *28*, 429-440.

Roberts, J. H., & Lattin, J. M. (1997). Consideration: Review of research and prospects for future insights. *Journal of Marketing Research*, *34*, 406-410.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15-22.

Saks, A. M., & Belcourt, M. (2006). An investigation of training activities and transfer of training in organizations. *Human Resource Management*, *45*(4), 629-648.

Sanderson, P. H. (1973). The 'don't know' option in mcq examinations. *British Journal of Medical Education*, *7*, 25-29.

Schuwirth, L. W. T., & Vleuten, C. P. M. van der. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, *38*, 974-979.

Schwaninger, A. (2003a). Training of airport security screeners. *AIRPORT 05/2003*, 11-13.

Schwaninger, A. (2003b). Evaluation and selection of airport security screeners. *AIRPORT*, *2*, 14–15.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. , 31–36.

Schwaninger, A. (2005a). Praxisfelder der wahrnehmungspsychologie. In B. K. und M. Groner (Ed.), (chap. Object recognition and signal detection). Bern: Huber.

Schwaninger, A. (2005b). Increasing efficiency in airport security screening. In *WIT Transactions on the Built Environment* (Vol. 82, pp. 407–416).

Schwaninger, A. (2005c). X-ray imagery: enhancing the value of the pixels. *Aviation Security International*, *Oct*, 16-21.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. In *IEEE ICCST Proceedings* (Vol. 38, pp. 258–264).

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. In *IEEE Aerospace and Electronic Systems* (Vol. 20(6), pp. 29–35).

Schwaninger, A., & Hofer, F. (2004). The internet society 2004, advances in learning, commerce and security. In K. Morgan & M. J. Spector (Eds.), (chap. Evaluation of CBT for increasing threat detection performance in X-ray screening). Wessex: WIT Press.

Schwaninger, A., Hofer, F., & Wetter, O. E. (2007). Adaptive computer-based training increases on the job performance of x-ray screeners. In *Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 8-11, 2007* (p. 117-124).

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements. In *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization, ACM Press, New York, USA* (pp. 123–130).

Sherriffs, A. B., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, *45*, 81-90.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist*, *36*(10), 1138-1146.

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General*, *134*(4), 443-460.

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Secific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1171-1185.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34-50.

Spitz, G., & Drury, C. G. (1978). Inspection of sheet materials - test of model predictions. *Human Factors*, *20*, 521-528.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Stark, J. A. (2000). Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions on Image Processing*, *9*(5), 889-896.

Stryker, M. P. (1991). Temporal associations. *Nature*, *354*, 108-109.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, *2*, 55-82.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple views? comment on biederman and gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1494–1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Object recognition in man, monkey and machine. In M. J. Tarr & H. H. Bülthoff (Eds.), *Object recognition in man, monkey and machine* (pp. 1–20). Cambridge, MA: MIT Press.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependene in shape-recognition. *Cognitive Psychology*, *21*(2), 233-282.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *353*, 1295-1306.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 13, p. 992-1006).

Verfaillie, K. (1992). Variant points of view on viewpoint invariance. *Canadian Journal of Psychology*, *46*, 215-235.

Votaw, D. F. (1936). The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. *Journal of Educational Psychology*, *27*, 698-703.

Wallis, G. M., & Bülthoff, H. H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, *3*, 22-31.

Wang, G., Obama, S., Yamashita, W., Sugihara, T., & Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience*, *8*(12), 1768-1775.

Wang, M. J. J., & Drury, C. G. (1989). A method of evaluating inspector's performance differences and job requirements. *Applied Ergonomics*, *20*(3), 181-190.

Wang, M. J. J., Lin, S. C., & Drury, C. G. (1997). Training for the strategy in visual search. *International Journal of Industrial Ergonomics*, *20*, 101-108.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*(1), 1-11.

Waters, B. K. (1976). The measurement of partial knowledge: A comparison between two empirical option-weighting methods and rights-only scoring. *Journal of Educational Research*, *69*(7), 256-260.

Wesman, A. G. (1971). Educational measurement. In R. L. Thorndike (Ed.), (2nd ed., p. 81-130). Washington, DC: American Council on Education.

Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202-238.

Zimmerman, D. W., & Williams, R. H. (1982). Element of cahnce and comparative reliability of matching tests and multiple-choice tests. *Psychological Reports*, *50*, 975-980.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357-371.

Zimmerman, D. W., Williams, R. H., & Symons, D. L. (1984). Empirical estimates of the comparative reliability of matching tests and multiple-choice tests. *Journal of Experimental Education*, *52*, 179-182.

Curriculum Vitae

PERSONAL DATA

Name	Koller Saskia Melanie, lic. phil.
Address	Erachfeldstrasse 1b 8180 Bülach
Telephone	+41 43 268 91 65
E-Mail	s.koller@psychologie.uzh.ch
Nationality	Switzerland
Date and place of birth	02.02.1979, Zurich
Marital status	single

WORK EXPERIENCE

Since January 2007	Research Assistant at University of Zurich, Department of Psychology, General Psychology (Cognition)
March 2005 - May 2006	Undergraduate research assistant at University of Zurich, Department of Psychology, General Psychology (Cognition)
August 2004 - September 2004	Internship at Kantonsspital Winterthur, Kinderklinik, Sozialpädiatrisches Zentrum
July 2002 - September 2002	Internship at Zurich State Police, Police Psychologist

November 1999 - June 2006	Part-time job Assistant Marketing/Sales, Cofex AG, Schwerzenbach (20-40%)

EDUCATION

Since January 2008	Doctoral student at University of Zurich, Department of Psychology, General Psychology (Cognition)
1999 - 2006	Studies of Psychology, Business Management and Criminology, University of Zurich (Lizenziat: December 2006)
1994 - 1999	High school, Kantonsschule Zürcher Unterland (Typus D)
1992 - 1994	Secondary school Glattfelden
1986 - 1992	Primary school Glattfelden

PERSONAL SKILLS

Languages

German	Mother tongue
English	Very good reading and writing
French	Good reading and writing
Italian	Good reading and writing

PUBLICATIONS

Peer reviewed articles

Koller, S.M., Hardmeier, D., Michel, S., & Schwaninger, A. (2008). Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-ray image interpretation. *Journal of Transportation Security, 1*(2), 81-106.

Michel, S., Koller, S.M., Ruh, M., & Schwaninger, A. (2007). Do "image enhancement" functions really enhance x-ray image interpretation? In D. S. McNamara & J.G. Trafton (Eds.),

Proceedings of the 29th Annual Cognitive Science Society (pp. 1301-1306). Austin, TX: Cognitive Science Society.

Koller, S.M., Hardmeier, D., Michel, S., & Schwaninger, A. (2007). Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation. In D. S. McNamara & J.G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1181-1186). Austin, TX: Cognitive Science Society.

Koller, S., & Schwaninger, A. (2006). Assessing x-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006,* Belgrade, Serbia and Montenegro, June 24-28, 2006, 393-397.

Conference papers

Michel, S., de Ruiter, J.C., Hogervorst, M., Koller, S.M., Moerland, R., & Schwaninger, A. (2007). Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners. *Proceedings of the 41st Carnahan Conference on Security Technology,* Ottawa, October 8-11, 2007, 201-207.

Michel, S., Koller, S., Ruh, M., & Schwaninger, A. (2006). The effect of image enhancement functions on x-ray detection performance. *Proceedings of the 4th International Security Technology Symposium*, Washington D.C., USA, November 27 - December 1, 2006, 434-439.

Presentations

Koller, S.M. (2008). Theoretical Tests for Screener Certification. ECAC/TAIEX Multilateral Workshop on Screeners' Certification, Brussels, Belgium, May 6, 2008.

Koller, S.M. (2007). Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation. The 10th Congress of the Swiss Society of Psychology, Zürich, Switzerland, September 13, 2007.

Posters

Koller, S., & Schwaninger, A. (2006). Assessing x-ray image interpretation competency of airport security screeners. Poster presented at the 4th LizentiandInnen- und Doktorierendenkongress (LiDoKo), Zürich, Switzerland, June 16, 2006.